UNIVERSIDADE FEDERAL DO PARANÁ

LEVISTON DA SILVEIRA

MONTAGEM E ANOTAÇÃO PARCIAL DA SEQUÊNCIA GENÔMICA DA BACTÉRIA DIAZOTRÓFICA *Azospirillum brasilense* FP2

CURITIBA

2012

LEVISTON DA SILVEIRA

MONTAGEM E ANOTAÇÃO PARCIAL DA SEQUÊNCIA GENÔMICA DA BACTÉRIA DIAZOTRÓFICA *Azospirillum brasilense* FP2

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de Mestre em Bioinformática.

Orientador:
Emanuel Maltempi de Souza, Dr.
Co-orientador:
Roberto Tadeu Raittz, Dr.

CURITIBA
2012

TERMO DE APROVAÇÃO

LEVISTON DA SILVEIRA

Montagem e anotação parcial da sequência genômica da bactéria diazotrófica *Azospirillum brasilense* FP2

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Dr. Emanuel Maltempi de Souza

Coorientador: Prof. Dr. Roberto Tadeu Raittz

Profª Drª Leda Satie Chubatsu
Universidade Federal do Paraná - UFPR

Prof. Dr. Marcelo Müller dos Santos
Universidade Federal do Paraná - UFPR

Curitiba, 28 de fevereiro de 2012

Dedico este trabalho

aos meus pais.

AGRADECIMENTOS

A Deus, por ter me concedido saúde, determinação e capacidade para concluir este trabalho.

Aos meus pais, pelo apoio e incentivo em todos os momentos, pela educação que me deram e por estarem sempre presentes nos momentos de dificuldades.

Aos meus orientadores, Prof. Dr. Emanuel Maltempi de Souza pela ajuda e orientação no direcionamento do trabalho e por compartilhar seu tempo e experiência profissional. Ao Prof. Dr. Roberto Tadeu Raittz pela ajuda em todas as etapas do trabalho, principalmente na área computacional, pelos questionamentos e conselhos durante o curso e pela amizade demonstrada.

A todos os colegas de curso de 2010 e 2011 sem exceção. Em especial a Vanely Souza e a Paula Saizaki por ter compartilhado todos os momentos de dificuldades, pela troca de experiências de trabalho e conhecimento em todo o período de execução da pesquisa.

Aos professores e colegas colaboradores, em especial ao Prof. Msc. Dieval Guizelini pela colaboração fundamental no desenvolvimento de ferramentas e pela discussão de soluções para melhoria dos resultados. Ao Prof. Dr. Lucas Ferrari de Oliveira pela disposição em ajudar e por ter cedido os servidores de seu projeto para tarefas relacionadas ao trabalho, ao Prof. Dr. Adriano Barbosa, ao doutorando Vinicius Weiss, ao Dr. Helisson Faoro e aos ICs Alysson, Gustavo e Ricardo pela ajuda e colaboração.

Ao Programa de Pós-Graduação em Bioinformática pela oportunidade concedida e em especial as secretárias Suzana e Lea pela disposição e empenho para ajudar sempre que precisei.

Aos órgãos financiadores CNPq, CAPES, REUNI e ao INCT pelo apoio financeiro fundamental para a minha dedicação integral ao trabalho.

A Universidade Federal do Paraná pela oportunidade concedida e por disponibilizar sua infra-estrutura física e humana para o desenvolvimento deste trabalho.

RESUMO

Bactérias fixadoras de nitrogênio ou diazotróficas são organismos capazes de reduzir o dinitrogênio atmosférico em amônia, um processo chamado fixação biológica de nitrogênio que é catalisado pela enzima nitrogenase. *Azospirillum brasilense* é uma bactéria fixadora capaz de se associar a raízes de gramíneas como trigo e arroz, e promover o crescimento do vegetal e aumento de produtividade. Neste trabalho utilizamos 5.768.466 milhões de seqüência de nucleotídeos obtidas em seqüenciador Solexa/Illumina e 112.782.028 milhões de seqüências de nucleotídeos obtidas em seqüenciador SOLID que foram utilizadas para obter uma seqüência parcial do genoma da espécie *Azospirillum brasilense* estirpe FP2. Para a montagem foi utilizado o montador Velvet para os dados Solexa e um pacote de programas adaptados para a montagem SOLID. Para finalização e correção da montagem, as montagens parciais de cada um dos métodos de seqüenciamentos foram mescladas, resultando em preenchimento de falhas. Como resultado final obtivemos uma sequência com 6.931.925 pb (386 scaffolds e 57 contigs) com aproximadamente 103.630 pb não determinadas. A sequência foi ordenada e alinhada ao genoma de *A. brasilense* estirpe Sp245 que foi utilizada como sequência de referência. A anotação parcial da sequência permitiu identificar 6119 genes, entre estes os envolvidos no metabolismo do nitrogênio.

Palavras-chave: *Azospirillum brasilense* FP2*,* fixação de nitrogênio, sequência genômica, montagem de genoma.

ABSTRACT

Nitrogen-fixing or diazotrophic bacteria are able to reduce atmospheric dinitrogen to ammonia, a process called biological nitrogen fixation which is catalyzed by the enzyme nitrogenase. *Azospirillum brasilense* is nitrogen-fixing bacteria that can be associated with roots of grasses such as wheat and rice, promoting plant growth and increasing productivity. In this work we used 5,768,466 million reads from a Solexa/Illumina sequencer and 112,782,028 million reads from a SOLID sequencer to assemble a partial genome sequence of *Azospirillum brasilense* FP2. The assembler Velvet was used for the Solexa data and a package of programs from Applied Biosystems was used to assemble the data from the SOLID sequencer. The two assemblies were merged to reduce the number of gaps. As a final result we obtained a sequence of 6,931,925 bp with approximately 103,630 bp not determined. The draft sequence was ordered and aligned to the genome of *A. brasilense* strain Sp245 which was used as reference. The annotation of the sequence identified 6119 genes, including those involved in nitrogen fixation and nitrogen metabolism.

Keywords: *Azospirillum brasilense* FP2*,* nitrogen fixation, genome sequence, genome assembly.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 *Azospirillum brasilense*

Bactérias do gênero *Azospirillum* são naturalmente encontradas na rizosfera de gramíneas cultiváveis tais como milho, trigo, sorgo e arroz (OKON e VANDERLEYDEN, 1997). Extensos estudos genéticos e bioquímicos sugerem que o Azospirillum seja um dos mais versáteis grupos de rizobactérias promotoras do crescimento de plantas (STEENHOUDT e VANDERLEYDEN, 2000). São 10 as espécies de Azospirillum: *Azospirillum lipoferum*, *Azospirillum brasilense* (TARRAND *et al.*, 1978), *Azospirillum amazonense* (MAGALHÃES *et al*., 1983), *Azospirillum halopraeferens* (REINHOLD *et al*., 1987), *Azospirillum irakense* (KHAMMAS, *et al*., 1989), *Azospirillum largimobile* (BEM DEKHIL *et al.*, 1997), *Azospirillum doebereinerae* (ECKERT *et al*., 2001), *Azospirillum oryzae* (XIE e YOKOTA, 2005), *Azospirillum melinis* (PENG *et al*., 2006) e *Azospirillum canadense* (MEHNAZ *et al*. 2007).

Bactérias deste gênero colonizam predominantemente a superfície de raízes, sendo que algumas estirpes são capazes de infectar e colonizar o interior dos tecidos de muitos cereais (DÖBEREINER *et al,.* 1995). Estudos mostraram que cerca de 1% do total de aeróbios heterotróficos encontrados em solos cultivados com arroz constitui de bactérias do gênero Azospirillum (LADHA *et al*., 1987).

Entre as bactérias do gênero *Azospirillum* a espécie *Azospirillum brasilense* está entre as mais estudadas. Estudos realizados com diferentes estirpes de *Azospirillum brasilense* mostraram que esta bactéria é capaz de estimular o crescimento da parte aérea e da raiz de vários cultivares de arroz de sequeiro (DIDONET *et al*., 2003).

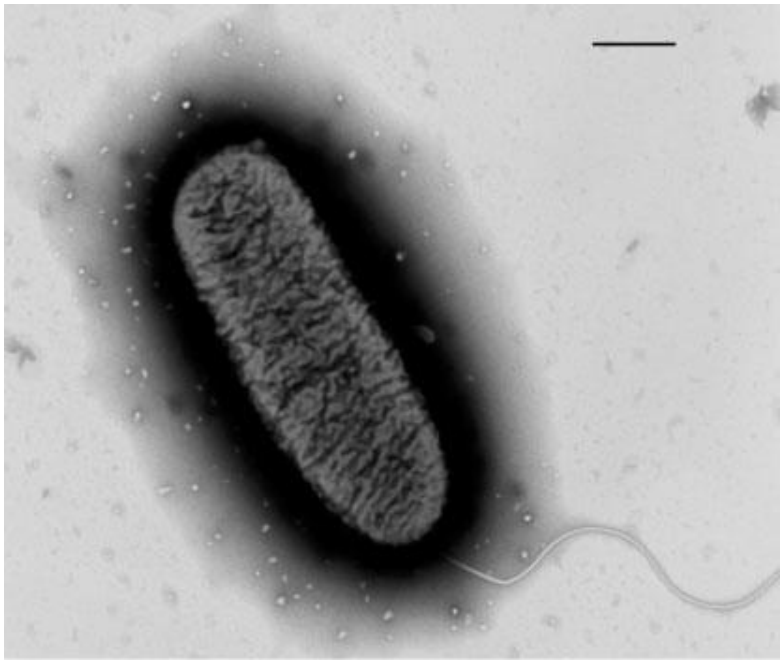

FIGURA 1 – MICROGRAFIA ELETRÔNICA MOSTRANDO A MORFOLOGIA DA CÉLULA ISOLADA. BARRA SUPERIOR: 0.6 µm.

FONTE: MEHNAZ *et al*., 2007.

Martin-Didonet *et al.* (2000) estudou a estrutura genômica de 10 estirpes, pertencentes a cinco espécies de *Azospirillum*, utilizando eletroforese de campo pulsado. Estes autores mostraram a presença de 16S rDNA em mais de um replicon, sugerindo que o *Azospirillum* possuí vários mega replicons variando de 0,2 à 2,7 Mpb. As estirpes FP2, SP7 e CD mostraram o mesmo perfil de DNA, replicons com tamanhos variando de 0,15 a 2,6 Mpb. A estirpe FP2 contém 7 replicons de tamanhos variados e com tamanho total do genoma estimado em 6,7 Mpb (TABELA 1). Apesar do papel destes replicons na ecologia e na sobrevivência dessas espécies ainda não ter sido determinado, é possivel que esta característica possa explicar a distribuição excepcional e a flexibilidade metabólica de membros deste gênero.

Em 2010, Kaneko *et al.* mostrou que o genoma completo da bactéria *Azospirillum sp.* estirpe B510 consiste de um único cromossomo (3 211 395 pb) e seis plasmídeos designados como pAB510a (1 455 109 pb), pAB510b (723 779 pb), pAB510c (681 723 pb), pAB510d (628 837 pb), pAB510e (537 299 pb), e pAB510f

(261 596 pb). O cromossomo tem 2893 genes e os plasmídeos um total de 3416 genes, totalizando 6309 regiões codificadoras de proteínas. Os resultados confirmaram os dados prévios de Martin-Didonet *et al.* (2000) que indicavam que os genomas de bactérias do gênero *Azospirillum* apresentam múltiplos megareplicons.

TABELA 1 – TAMANHO MOLECULAR DOS REPLICONS DE *Azospirillum spp.*

Especies	**Estirpe**	**Tamanho (Mpb)**	
		Replicons[d] em Mpb	Genoma Em Mbp (Estimado)
A. brasilense	FP2	2,5[a,b]; 1,72[a]; 0,81[a] (L); 0,7 (L); 0,63[a] (L); 0,17; 0,15	6,7
	Sp7	2,5[a,b]; 1,74[a]; 0,81[a] (L); 0,70 (L); 0,64[a] (L); 0,21; 0,2	6,8
	Cd	2.6[a,b]; 1.77[a]; 0.81[a] (L); 0.71 (L); 0.64[a] (L); 0.21; 0.19	6,9
	Sp245	2,6[a,b]; 1,76[a]; 0,9[a] (L); 0,78 (L); 0,72[a] (L); 0,21; 0,14	7,1
A. lipoferum	Sp59b	2,6[a]; 1,8[a]; 1,38[a]; 1,18[a] (L); 0,97[a] (L); 0.71 (L); 0.65[a] (L); 0.4	9,7
	JA25	2,25(ND); 1,8[a]; 1,1[a] (L); 0,85[a] (L); 0,55[a] (L); 0,45 (L); 0,3; 0,27; 0,22; 0,15	7,9
A. amazonense	Y2	2,7[a]; 2,2; 1,7[a]; 0,75	7,3
	Y6	2,6[a]; 2,1; 1,8[a]; 0,71	7,2
A. irakense		2,4[a]; 1,2; 0,95[a]; 0,22	4,8
A. halopraeferens		2,6[a]; 1,2; 0,98[a]; 0,92[a];0,22	5,9

[a]Hibridiza com *16S rDNA.*
[b]Hibridiza com *nifHDK.*
[c]Os tamanhos moleculares indicados são as médias de pelo menos cinco determinações (exceto para *A. irakense* e *A. halopraeferens*, para os quais 2 experimentos foram realizados).
[d]Abreviaturas: (L), indicação de molécula linear.

FONTE: MARTIN-DIDONET *et al*., 2000.

1.1.1 Promoção de crescimento vegetal

A capacidade de promoção de crescimento de plantas de alguns microorganismos permitem aumentar a produtividade fazendo melhor uso de nutrientes do solo e não necessitando assim de outros fertilizantes (KENNEDY *et al.* 2006). Experimentos de inoculação com espécies do genêro *Azospirillum* vem sendo realizados em diferentes países. Em vários desses ensaios se demonstrou aumento no conteúdo de nitrogênio, fósforo, potássio e outros minerais importantes para o crescimento da planta. Em cerca de 70 % destes ensaios comprovou-se o aumento de produtividade de até 30% (DOBEREINER e PEDROSA, 1987; LADHA, 2000).

A inoculação com espécies do gênero *Azospirillum* promovem incrementos significativos no desenvolvimento radicular das plantas, resultando no melhor aproveitamento e utilização de adubo e água e consequentemente, o melhor desenvolvimento das plantas (BALDANI *et al.* 1997). A adição de bactérias com esse potencial podem resultar em benefícios econômicos e ambientais como altos rendimentos, redução de custo de fertilizantes e baixa emissão de gases com efeito estufa (N_2O).

Para Summer (1990), bactérias diazotróficas isoladas da mesma variedade da planta que se deseja inocular, são mais eficientes. Organismos que são adaptados às condições ambientais da região podem apresentar melhores condições para concorrer com a microbiota nativa, tanto na relação planta-solo-clima (edafoclimática) como pelo aumento populacional promovido pela inoculação. Por tanto, existe a necessidade da seleção de microrganismos promissores capazes de fixar nitrogênio e produzir substâncias promotoras do crescimento de plantas para inoculação. Estudo apresentado por Roesch *et al*. (2005) com o objetivo de avaliar o suprimento de nitrogênio via fixação biológica e a capacidade de isolados de bactérias diazotróficas promover o aumento radicular da planta do trigo, mostrou que dois isolados de bactérias do gênero *Azospirillum* testados aumentaram 2 a 3 vezes o comprimento radicular comparado aos tratamentos sem inoculação (FIGURA 2).

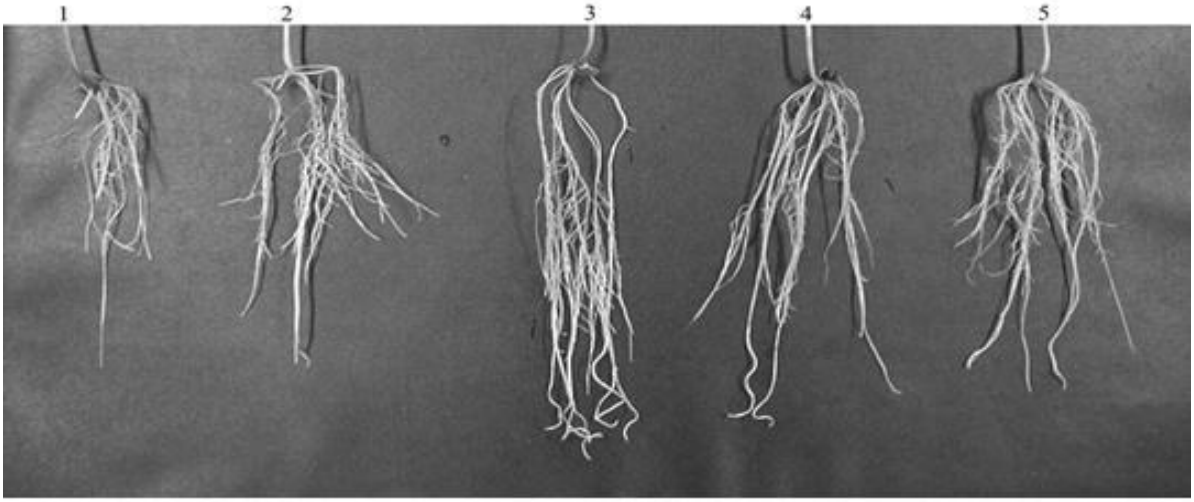

FIGURA 2 - PLANTA DE TRIGO 21 DIAS APÓS A GERMINAÇÃO. 1 - TAMANHO COM ADUBAÇÃO NITROGENADA E SEM INOCULAÇÃO; 2 - TRATAMENTO SEM ADUBAÇÃO NITROGENADA E SEM INOCULAÇÃO; 3 - TRATAMENTO COM INOCULAÇÃO DO ISOLADO Nº 1; 4 - TRATAMENTO COMO INOCULAÇÃO DO ISOLADO Nº 2; 5 - TRATAMENTO COM INOCULAÇÃO DE UM ISOLADO DO GÊNERO AZOSPIRILLUM.

FONTE: ROESCH, *et al*., 2005.

Para Kennedy *et al.* (2006), a efetiva promoção do crescimento de plantas através da inoculação com rizobactérias exigirá estratégias específicas para fazer uso de todos os fatores benéficos. O estudo do genoma de bactérias com esse potencial biotecnológico (promoção do crescimento de plantas), permitirá uma melhor compreensão dos efeitos benéficos para a planta associada, mecanismos de promoção de crescimento e da ecologia desta bactéria.

1.2 O SEQUENCIAMENTO DE GENOMAS

Já se passaram mais de 50 anos desde a descoberta de que o DNA é responsável por armazenar as informações genéticas e a partir de então se iniciou uma busca por uma forma de se obter e decodificar a informação localizada nos cromossomos (MIR, 2004). O seqüenciamento rápido de genomas tem sido um dos grandes desafios da genômica, ciência que estuda a estrutura e funcionamento do material genético de uma espécie (CHAN, 2005).

Enormes progressos têm sido feito nos últimos 15 anos na obtenção de seqüências de um genoma completo. Desde o seqüenciamento dos primeiros

genomas completos (*Haemophilus influenza* e *Mycoplasma genitalium)*. O surgimento de novas tecnologias de seqüenciamento denominadas seqüenciamento de nova geração nos últimos 7 anos, tem contribuído para que esse número de informações aumente cada vez mais (CHAN, 2005; MIR, 2004). O estudo destas informações nos permite entender aspectos integrados da biologia desses organismos, inter-relacionar seqüência, estrutura tridimensional, padrões de expressão, interações e função de proteínas individuais e complexos proteínas-ácidos nucléicos, entender a história evolucionária e nortear modificações científicas de sistemas biológicos e apoiar aplicações nas áreas de medicina, agricultura e tecnologia (LESK, 2008).

1.2.1 Tecnologias de sequenciamento

Tecnologias de seqüenciamento podem ser classificadas como o conjunto de instrumentos, materiais, protocolos e métodos que estão envolvidos na coleta, preparação e isolamento da amostra para o seqüenciamento e a montagem da sequência final (CHAN, 2005). Um longo caminho é percorrido até o processo final de montagem e análise da seqüência, normalmente dividida em duas fases relacionadas: uma fase experimental de bancada e uma fase de análise computacional (FIGURA 3). As novas tecnologias de seqüenciamento evoluem rapidamente e tem a vantagem de gerar informação sobre milhões de pares de bases em uma única corrida.

Dentre as várias tecnologias que estão surgindo, podemos citar a plataforma 454 FLX da Roche, a Solexa da Illumina e a plataforma da Applied Biosystems, denominada SOLID System, além de outras que começam a ser utilizadas como a HeliscopeTrue Single Molecule Sequencing (tSMS), da Helicos (CARVALHO e SILVA, 2010).

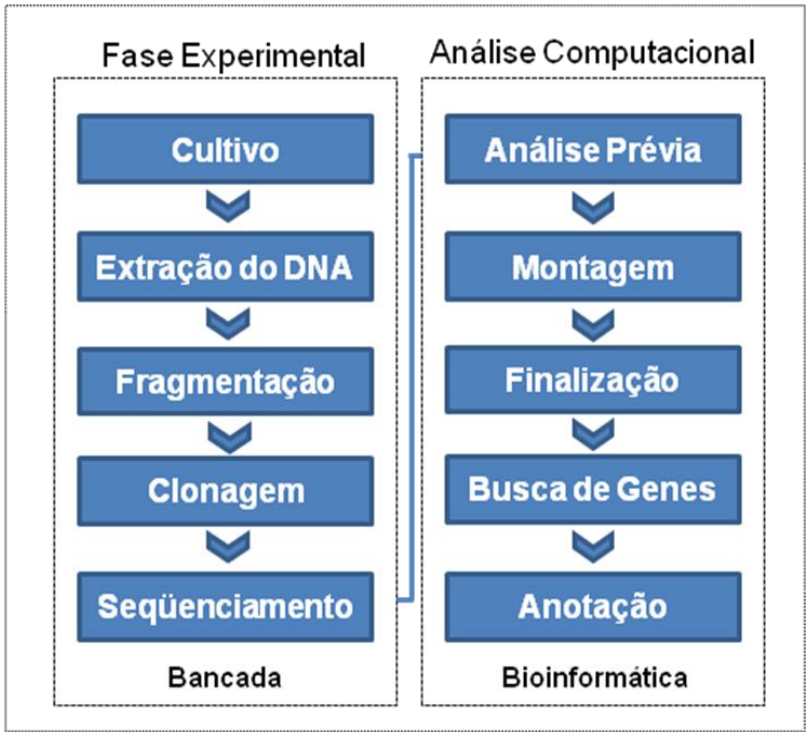

FIGURA 3 - ETAPAS DE UM PROCESSO DE SEQUENCIAMENTO E FINALIZAÇÃO. A FASE EXPERIMENTAL SE CARACTERIZA POR EXPERIMENTOS EM BANCADA (EXPERIMENTOS BIOLÓGICOS) E A SEGUNDA FASE POR ANÁLISE COMPUTACIONAL (LABORATÓRIO DE BIOINFORMÁTICA).

1.2.1.1 Plataforma Illumina

Nesta plataforma, como no método de Sanger, o sequenciamento (SANGER, *et al.*, 1977) é realizado por síntese usando DNA polimerase e nucleotídeos marcados com diferentes fluoróforos. A grande inovação desta técnica é a clonagem dos fragmentos *in vitro* diretamente em uma superfície sólida de vidro, superfície de clonagem, processo conhecido como PCR de fase sólida (FEDURCO *et al*., 2006; TURCATTI *et al*., 2008). A superfície consiste de uma lâmina de vidro dividida em oito linhas onde as extremidades dos fragmentos são ligados.

No processo de seqüenciamento os fragmentos de DNA têm suas extremidades (5’ e 3’) ligadas a adaptadores (FIGURA 4-A) e fixados a superfície de clonagem, de modo que a extremidade 3’ fique sempre livre para servir na iniciação da reação de seqüenciamento dos fragmentos imobilizados no suporte (FIGURA 4-B).

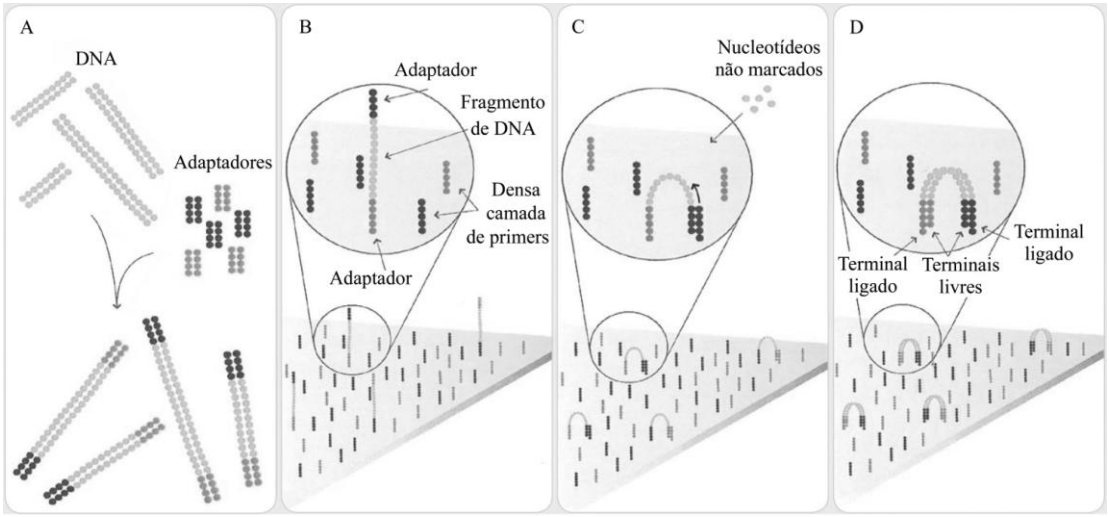

FIGURA 4 - PROCESSO INICIAL DE SEQUENCIAMENTO DA PLATAFORMA ILLUMINA.

FONTE: CARVALHO e SILVA, 2010.

A grande quantidade de adaptadores fixos na superfície de clonagem facilita a hibridização do adaptador livre (3') e sua seqüência complementar fixa, próximo ao clone inicial durante o ciclo de anelamento. No ciclo de anelamento o fragmento forma uma estrutura em "ponte" na superfície de seqüenciamento (FIGURA 4 - C e D) e após o fornecimento de reagentes necessários a extensão ocorre (PCR), formando a fita complementar também em "ponte". (CARVALHO e SILVA, 2010).

No chamado ciclo de desnaturação (por elevação da temperatura) as fitas são separadas e linearizadas (FIGURA 5 - E), o ciclo (de anelamento) é repetido por 35 vezes (FIGURA 5 - F) fazendo com que as mil cópias geradas de cada fragmento se agrupem formando um *cluster* de seqüenciamento (FIGURA 5 - G). Após esta etapa, nucleotídeos terminadores marcados são fornecidos para as reações que ocorrem dentro do *cluster* (FIGURA 5 – H).

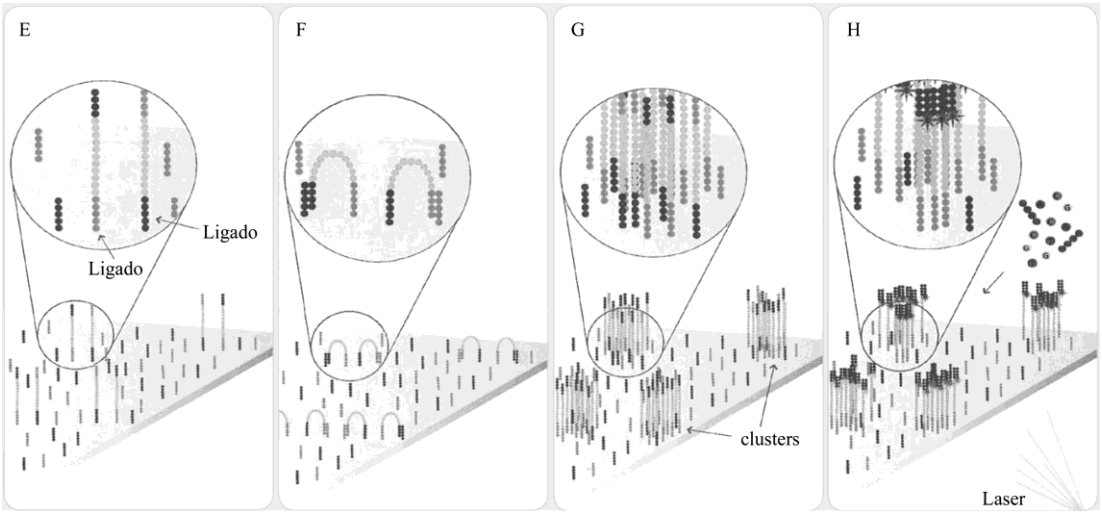

FIGURA 5 - PROCESSO INTERMEDIÁRIO DE SEQUENCIAMENTO DA PLATAFORMA ILLUMINA.

FONTE: CARVALHO e SILVA, 2010.

A alta densidade de *clusters* formada possibilita que o sinal de fluorescência gerado com a incorporação de cada um dos nucleotídeos terminadores tenha uma intensidade suficiente para a detecção confiável (CARVALHO e SILVA, 2010). Na etapa seguinte, com a incorporação de nucleotídeos terminadores marcados e excitação a laser (FIGURA 5 - H), é gerado um sinal que é captado por dispositivos de leitura e interpretado como um dos quatro possíveis nucleotídeos marcados (FIGURA 6 - I, J, K), e o processo é repetido para cada nucleotídeo que compõe a seqüência. A leitura das bases é feita pela análise seqüencial das imagens capturadas (FIGURA 5 - L) em cada ciclo de seqüenciamento. Em geral, leituras de 25-35 bases são obtidas em cada *cluster* (SHENDURE e JI, 2008).

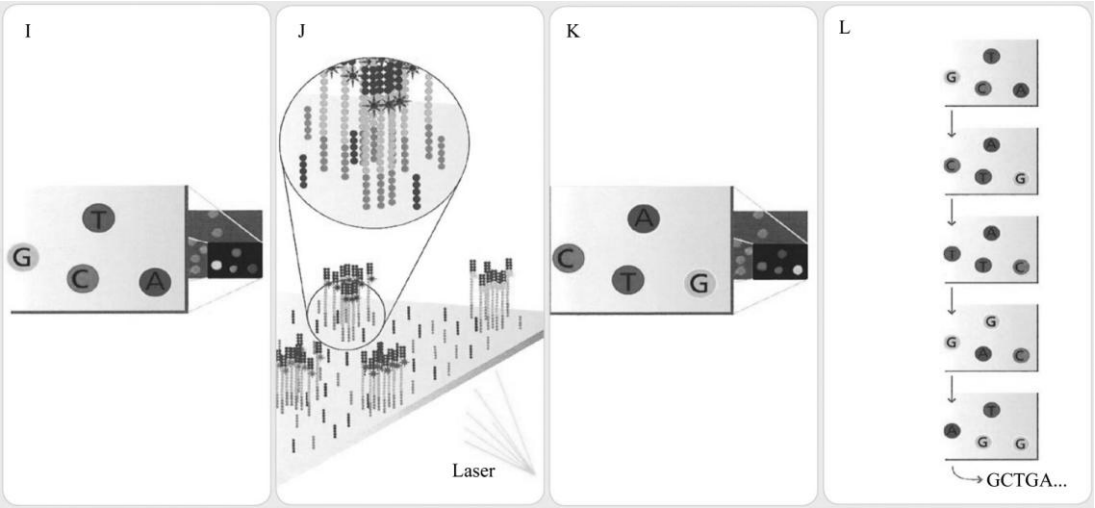

FIGURA 6 - PROCESSO FINAL DE SEQUENCIAMENTO DA PLATAFORMA ILLUMINA.

FONTE: CARVALHO e SILVA, 2010

1.2.1.2 Plataforma SOLID

Na plataforma SOLID (MCKERNAN *et al*., 2006) a reação de catálise acontece por ação de uma DNA ligase, e não uma polimerase. Após a fragmentação mecânica do DNA por um sonicador (60-90 pb), para bibliotecas de fragmentos, ou 1-10 Kb, para biblioteca pareada ou *"mate-pair",* os fragmentos são ligados a adaptadores universais (P1 e P2) em ambas as extremidades (FIGURA 7).

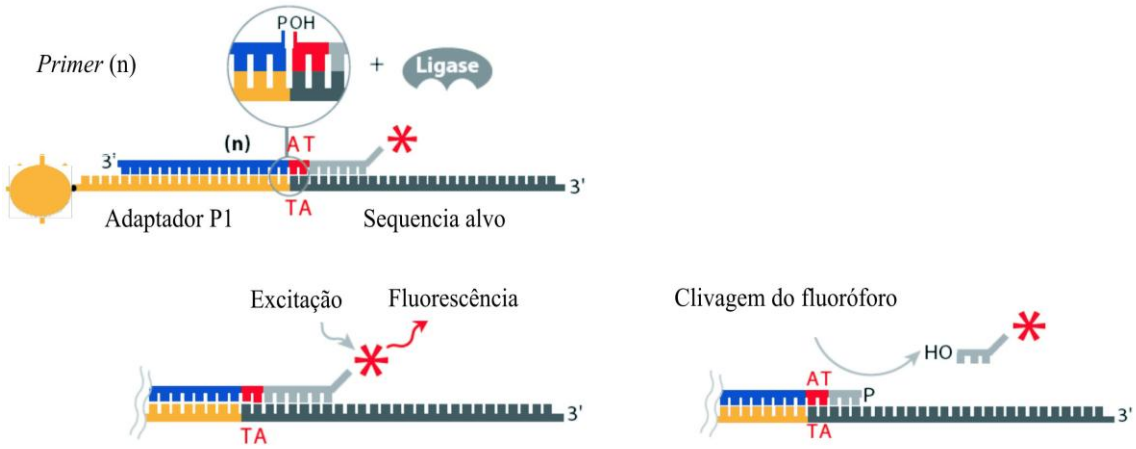

FIGURA 7 – PREPARO DE BIBLIOTECA PARA SEQUENCIAMENTO.

Em seguida a biblioteca é amplificada e, após seleção de tamanhos, os fragmentos são ligados por hibridização com seqüências complementares a adaptadores fixos a microesferas metálicas e amplificadas novamente em uma PCR de emulsão. Na primeira etapa do seqüenciamento os moldes ligados as esferas são adicionados aos primers (n), a enzima ligase e as sondas (FIGURA 7). O seqüenciamento é dividido em etapas distintas com o uso de primer universal com n bases na primeira etapa, n-1 na segunda, e assim sucessivamente até a última etapa, onde o primer é n-4 bases (CARVALHO e SILVA, 2010). A cada etapa um dinucleotídeo do DNA molde é determinado através da ligação de sonda que é identificada pelo sinal de fluorescência. A sonda contém 8 bases, sendo que duas correspondentes ao dinucleotídeo específico, as bases 6, 7 e 8 carregam o fluoróforo marcador e as demais bases são degeneradas em todas as combinações possíveis. As três últimas bases da sonda são removidas (FIGURA 7 e 8) e um novo ciclo de hibridização se repete até que o alvo seja inteiramente coberto (35pb).

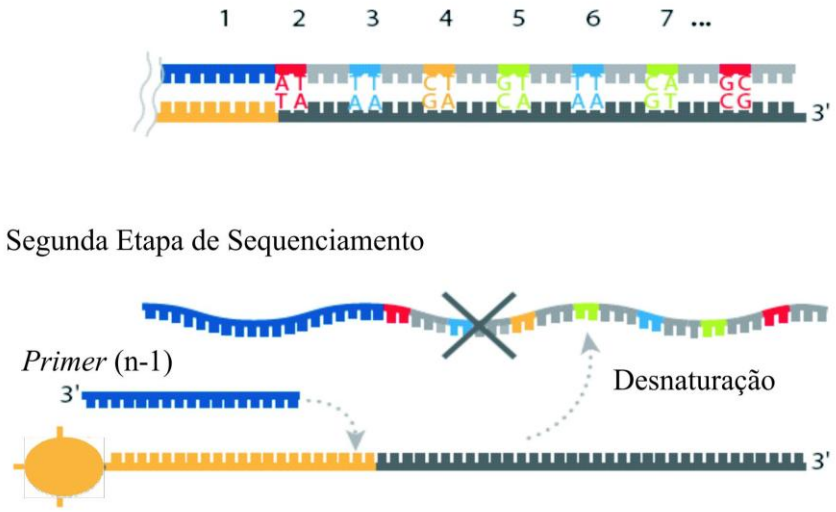

FIGURA 8 – CICLO DE HIBRIDIZAÇÃO.

FONTE: CARVALHO e SILVA, 2010

Ao final da primeira etapa, a fita dupla é desnaturada, e uma nova etapa de seqüenciamento recomeça com a utilização de um primer n-1, seguida de etapas com, primers n-2, n-3 e n-4 de forma que toda a seqüência alvo seja determinada (FIGURA 9). Para que a sequência do dinucleotídeo da sonda seja resolvida são

necessárias leituras (a primeira e a segunda) de duas etapas (CARVALHO e SILVA, 2010).

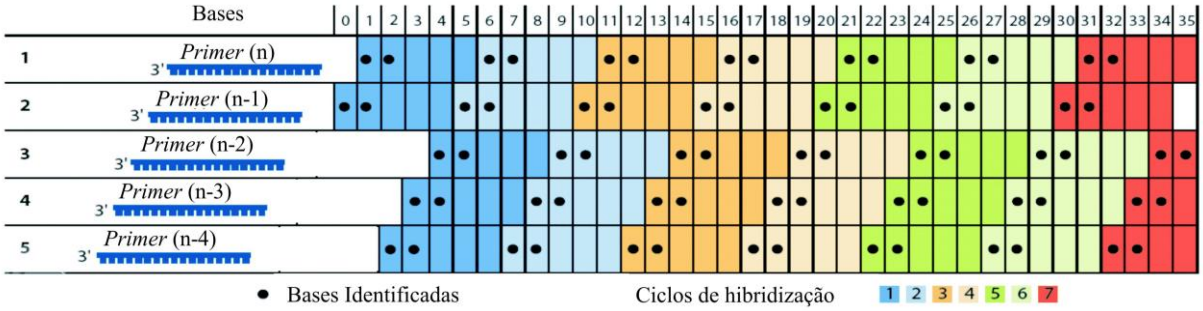

FIGURA 9 – CINCO ETAPAS COM SETE CICLOS DE HIBRIDIZAÇÃO.

FONTE: CARVALHO e SILVA, 2010.

A decodificação de sinal de leitura é feita com base na combinação dos sinais fluorescentes. Cada sinal de fluorescência especifica um dinucleotídeo e não uma única base. Como as bases do adaptador P1 são conhecidas, a identificação da primeira base e da segunda etapa é possível quando se utiliza o primer n-1. Os demais sinais são especificados pela combinação única de cores a partir da base conhecida (FIGURA 10).

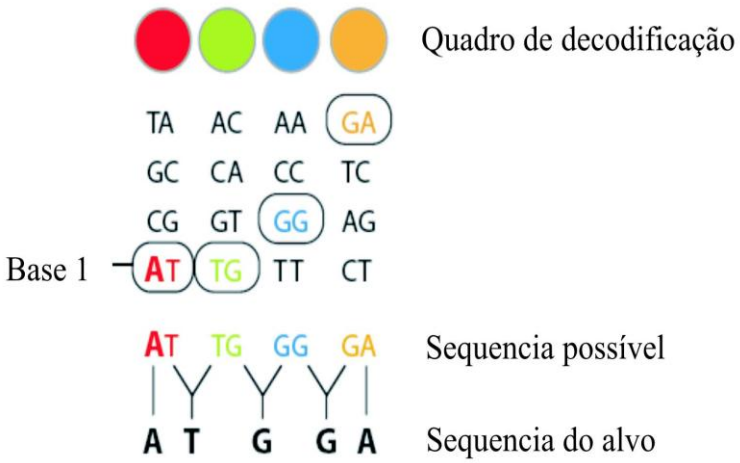

FIGURA 10 – DINUCLEOTÍDEOS POSSÍVEIS PARA CADA SINAL DE COR.

FONTE: CARVALHO e SILVA, 2010.

As leituras curtas resultantes do seqüenciamento SOLID são normalmente utilizadas para reconstrução da sequência do genoma de bactérias e com bibliotecas em mate-paired (DURFEE *et al*., 2008). A grande quantidade de dados gerados tornou a plataforma SOLID especialmente útil no estudo de transcriptoma (PASSALACQUA *et al*., 2009)

1.3 MONTAGEM E ANOTAÇÃO DE GENOMAS

A seqüência do genoma do organismo é rica em informações biológicas, e deve ser anotada através de análises computacionais e interpretação biológica para obter a maior quantidade possível de dados úteis (LEWIS *et al*., 2000). Quando se possui a informação completa do genoma do organismo é possível se determinar a localização física dos genes, RNAs, elementos repetitivos, etc. (ROUZÉ, 1999). Quando o foco é a anotação de proteínas, o alvo é inferir a sua função biológica a partir de comparação com seqüências de função conhecida (PROSCOCIMI, 2007).

1.3.1 Montagem de genomas

Atualmente são utilizadas duas principais estratégias de seqüenciamento, o direcionado e o aleatório, que se diferenciam apenas no processo de clonagem, tamanho de insertos e na maneira de seqüenciamento (STERKY e LUNDEBERG, 2000). Seja qual for à técnica escolhida o DNA deve ser fragmentado, assegurando que nenhuma região do genoma ficará sem representação, ou seja, que no processo de montagem haja sobreposição suficiente dos fragmentos para uma montagem completa.

Um método muito utilizado para seqüenciamento de genoma de procariotos é conhecido como *shotgun* (seqüenciamento aleatório), onde o DNA é submetido a um processo que fragmenta as moléculas (MEIDANIS e SETÚBAL, 1994), utilizando, por exemplo, enzimas de restrição ou sonicação (STERKY e LUNDEBERG, 2000). Como resultado dessa fragmentação têm milhares de pequenos pedaços que são seqüenciados e então as seqüências são agrupadas gerando os contigs. O processo de montagem é importante por ainda não existir uma técnica de seqüenciamento que permita o seqüenciamento de moléculas de DNA com mais de mil pares de bases (PROSDOCIMI, 2007).

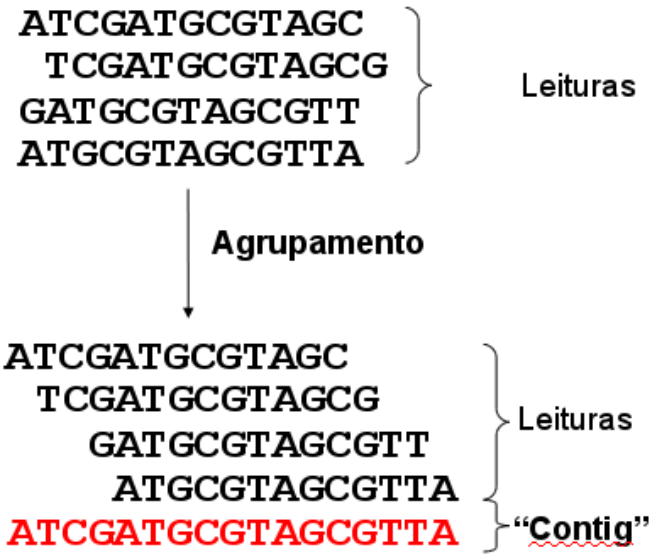

FIGURA 11 - MONTAGEM DAS LEITURAS. CONTIG É GERADO A PARTIR DO AGRUPAMENTO DE LEITURAS SOBREPOSTAS.

FONTE: RAMOS, 2011.

A maioria dos algoritmos de montagem segue um padrão complexo para obtenção de uma montagem completa do genoma (CARVALHO e SILVA, 2007). Montar um genoma consiste basicamente em agrupar as seqüências que foram geradas pelo seqüenciador, levando em consideração as regiões que são iguais entre essas seqüências (leituras), de forma a estendê-la por meio de sobreposição para reconstruir a sequência original (SHUSTER, 2008). A sequência obtida do processo de seqüenciamento é conhecida como leitura (reads). Quando duas ou mais dessas leituras são alinhadas por possuírem a mesma sequência de bases, são geradas sequências maiores, chamados de “contigs” (seqüências contínuas geradas pela sobreposição das leituras) (FIGURA 11). Os “contigs” podem ser alinhados gerando seqüências ainda maiores chamados de “scaffolds” (“contigs” agrupados), que na verdade são “supercontigs”.

1.3.2 Anotação de genomas

A anotação da sequência genômica consiste de procedimentos para se identificar genes e suas funções, por exemplo, identificação de sequência que correspondam a genes codificadores de proteínas, tRNA, rRNA (PROSDOCIMI, 2007).

FIGURA 12 – ESQUEMA REPRESENTANDO AS FASES DE ANOTAÇÃO.

FONTE: PROSDOCIMI *et al*, 2007.

Na anotação de genes codificadores de proteínas, procura-se comparar com genes já descobertos e caracterizar para descobrir suas funções e na anotação de processos procura identificar as vias metabólicas de processos biológicos nos quais diferentes genes interagem (FIGURA 13). Uma forma comum de se realizar anotação de proteínas é a busca de similaridade em diferentes bancos de dados com ferramentas de alinhamento local como BLASTp ou PSI-BLAST (ALTSCHUL *et. al*., 1990).

2 OBJETIVOS

2.1 Objetivo geral

Obter a sequência parcial do genoma da bactéria diazotrófica *Azospirillum brasilense* estirpe FP2, utilizando leituras curtas provenientes de seqüenciamento de diferentes tecnologias (Ilumina e Solid) e identificar regiões codificadoras de proteínas na sequência genômica obtida.

2.2 Objetivos específicos

- Montar o genoma da bactéria *A. brasilense* FP2 a partir de dados de seqüenciamento Solid e Ilumina;
- Ordenar os contigs e *scaffolds* obtidos tendo como referência a sequência genômica da estirpe *A. brasilense* Sp245
- Determinar lacunas (*gaps*) de montagem e definir estratégias de correções em falhas internas nos scaffolds.
- Definir estratégia para união de *scaffolds* a partir de montagens alternativas
- Identificar genes na seqüência parcial obtida a partir de anotação automática.

3 MATERIAIS E MÉTODOS

3.1 ORIGEM DOS DADOS

Para este trabalho foram utilizados dois conjuntos de dados de seqüenciamento de tecnologias distintas. Em ambas as tecnologias as duas extremidades dos insertos das bibliotecas foram sequenciadas e por isto neste trabalho as sequências obtidas foram denominadas “pareadas”. As seqüências obtidas com o Solid tinham inserto médio de 1,5 Kbp (biblioteca tipo *mate-pair*) enquanto as do sequenciador Illumina cerca de 0,25 Kpb (biblioteca tipo *pair-end*). O primeiro conjunto de dados consistiu de seqüências utilizando o seqüenciador **Illumina** e foi realizada pela empresa FASTERIS (Genebra, Suíça).

TABELA 2 – RESUMOS DOS DADOS DE SEQUENCIAMENTO OBTIDO COM O SEQUENCIADOR ILLUMINA

Tipo de leitura	*Mate-pair*
Tamanho das leituras	38 pb
Distância entre os pares	~300 pb
Total de leituras	5.768.466
Total de Bases	219.201.708 pb
Cobertura estimada do genoma	32x

O segundo conjunto de dados foi obtido com o sequenciador **Solid** (**Life Technologies)**, sendo obtido pelo Núcleo de Fixação Biológica de Nitrogênio no Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná.

TABELA 3 – RESUMO DOS DADOS OBTIDOS COM SEQUENCIADOR SOLID

Tipo de leitura	*Mate-pair*
Tamanho das leituras	50 pb
Distância entre os pares	~1500 – 2000 pb
Total de leituras	112.782.028
Total de Bases	5.639.101.400 pb
Cobertura estimada do genoma	782x

3.1.1 Genoma de referência

O genoma de referência utilizado foi da bactéria *Azospirillum brasilense* estirpe Sp245. A espécie selecionada não apresenta o genoma totalmente fechado (WISNIEWSKI-DYÉ *et al*., 2011)

TABELA 4 – GENOMA DE REFERÊNCIA PARA VALIDAÇÃO DOS DADOS

Referência GeneBank	**Identificação**	**Tamanho**
HE5732	*Azospirillum brasilense* Sp245 (67 contigs).	7,5 Mpb

FONTE: WISNIEWSKI-DYÉ *et al*., 2011

3.2 CONFIGURAÇÃO DE SISTEMAS

3.2.1 Servidores

Os servidores utilizados (TABELA 5) estão lotados no Laboratório de Bioinformática do Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná. Os computadores (TABELA 5) foram adquiridos com recurso do projeto Jovem Pesquisador (*CNPq Número 567035/2008-5,* coordenado pelo Prof. Lucas Ferrari).

TABELA 5 - SERVIDORES UTILIZADOS.

Sistema operacional	GNU/Linux Debian 6.0
Processador	AMD Phenom II X6 (1090t) Black Edition
Disco Rígido	1,5 TeraBytes
Memória RAM	16 GigaBytes
Placa de Vídeo	nVidia GeForce 9600 GT (1GigaByte)
Sistema operacional	GNU/Linux Debian 6.0
Processador	AMD Phenom X4 (955) Black Edition
Disco Rígido	1TeraBytes
Memória RAM	8 GigaBytes
Placa de Vídeo	nVidia GeForce 9600 GT (1GB)

Foi também utilizado um *cluster* SGI (TABELA 6) do Programa de Pós-Graduação em Bioinformática da Universidade Federal do Paraná. Este *cluster* está lotado no Departamento de Informática da UFPR.

TABELA 6 - *CLUSTER* UTILIZADO NO PROCESSO DE MONTAGEM

Sistema operacional	GNU/Linux Debian 6.0
Processador	64 Núcleos (Intel® Xeon® CPU E7 8837 @ 2.67GHz
Rede/Armazenamento	7.2 TeraBytes
Disco Rígido	200 GigaBytes
Memória RAM	512 GigaBytes
Arquitetura	NUMAlink 5

3.2.2 Computadores de mesa

Os computadores do Laboratório de Bioinformática (LABINFO) do Programa de Pós-Graduação em Bioinformática no Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná (TABELA 7) foram utilizados durante o desenvolvimento do projeto em várias atividades onde a demanda de processamento foi menor (alinhamento de seqüências, análise de dados de montagem, correção de falhas de montagem). Foi utilizado também o computador pessoal do mestrando (TABELA 8) para pequenas tarefas computacionais.

TABELA 7 - COMPUTADORES UTILIZADOS PARA EXECUÇÃO DO PROJETO

Sistema operacional	GNU / Linux Ubuntu 11.10
Processador	Intel® Core™ i5 CPU 650 @ 3.20GHz x 4
Disco Rígido	320 GigaBytes
Memória RAM	16 GigaBytes
Tipo de Sistema	64 bits
Sistema operacional	Windows Vista™ Home Basic
Processador	Intel® Pentium® Dual CPU E2200 @ 2.20GHz
Disco Rígido	227 GigaBytes
Memória RAM	2 GigaBytes
Tipo de Sistema	32 Bits

TABELA 8 - COMPUTADOR PESSOAL PARA EXECUÇÃO DE ATIVIDADES REMOTAS

Sistema operacional	GNU / Linux Ubuntu 11.10
Processador	Pentium® Dual-Core CPU E5700 @ 3.00GHz x 2
Disco Rígido	1 TeraByte
Memória RAM	4 GigaBytes
Placa de Vídeo	Intel® G41

3.3 PROGRAMAS UTILIZADOS

3.3.1 FastQC

O programa FastQC permitiu extrair relatórios dos conjuntos de sequências a partir dos dados brutos.

FastQ é um formato de arquivo gerado pelo seqüenciador Illumina que contêm a seqüência de nucleotídeos e juntamente a representação dos níveis de qualidade.

O FastQC suporta arquivos nos seguintes formatos FastQ

- Colorspace FastQ;
- GZip comprimido FastQ;
- SAM;
- BAM;
- SAM/BAM apenas de mapeamentos (normalmente usado para dados em color space).

Este programa permitiu gerar gráficos da distribuição das seqüências levando em consideração a qualidade média total por base, a qualidade média de todas as leituras e o conteúdo G + C dos dados (BABRAHAM BIOINFORMATICS).

3.3.2 Quality assessment

O programa Quality Assessment foi desenvolvido para filtrar e gerar gráficos que mostram a distribuição de valores de qualidade a partir de dados de seqüenciamento. Ele permite além da análise dos dados, aplicar filtro de qualidade aos dados (RAMOS *et al.,* 2011).

O Quality Assessment recebe como entrada dois arquivos: um arquivo com valores de qualidade phred para cada base de uma leitura e as seqüências de nucleotídeos ou em formato color space (Solid).

3.3.2.1 Avaliação da qualidade das sequências

A confiabilidade da identificação de cada base seqüenciada é dada pelo valor ou escore de qualidade da base. A métrica mais utilizada é o valor de qualidade Phred (Q) que é dado pela seguinte fórmula, onde P é a probabilidade da base ter sido errôneamente identificada (EWING *et al*., 1998)

$$Q = -10 \log_{10} P$$

TABELA 9 – QUALIDADE PHED RELACIONADA COM A PROBABILIDADE DE ERRO E PRECISÃO DA BASE.

Qualidade Phred	Probabilidade de Erro	Precisão
10	1 em 10	90%
20	1 em 100	99%
30	1 em 1000	99,9%
40	1 em 10000	99,99%
50	1 em 100000	99,999%

FONTE: Ramos, 2011.

As bases da extremidade 3’ nos seqüenciadores de nova geração (NGS) normalmente apresentam baixa qualidade (CHOU *et al.,* 2001), o que também ocorre com o método de Sanger manual ou automatizado (SANGER, *et al.*, 1977). As regiões da seqüência com baixo valor Phred, normalmente era eliminada para evitar erros de montagem. Selecionar as regiões com baixa probabilidade de erro de identificação de base melhora os alinhamentos por possibilitar a redução do valor atribuído a penalizações nos alinhamento (SMITH *et al.,* 2008; RAMOS, 2011).

3.3.3 Velvet

O programa Velvet é um conjunto de algoritmos criados para manipulação de grafo de Bruijn usados para montagem de seqüências genômicas (ZERBINO e BIRNEY, 2008). Um grafo de Bruijn é uma representação baseada em palavras curtas denominadas de k-mers, especialmente utilizado na reconstrução de

seqüências genômicas a partir de um conjunto de dados de leituras curtas (25-50pb). Com o emparelhamento e sobreposição de leituras curtas com a utilização do Velvet pode-se produzir contigs de tamanhos consideráveis (ZERBINO E BIRNEY, 2008).

No grafo de Bruijn cada nó N representa uma série de sobreposições de k-mers (palavras curtas) adjacente a sobreposição de k-mers por k-1 nucleotídeos. As informações contidas por um k-mer é o seu último nucleotídeo, sendo denominada de seqüências do nó, ou s (N). Nesses nós e arcos, as leituras são mapeadas como “caminhos” que atravessam o grafo. Extrair a seqüência de nucleotídeos de um caminho é relativamente simples, dado o k-mer inicial do primeiro nó e as seqüências de todos os nós no caminho (ZERBINO e BIRNEY, 2008).

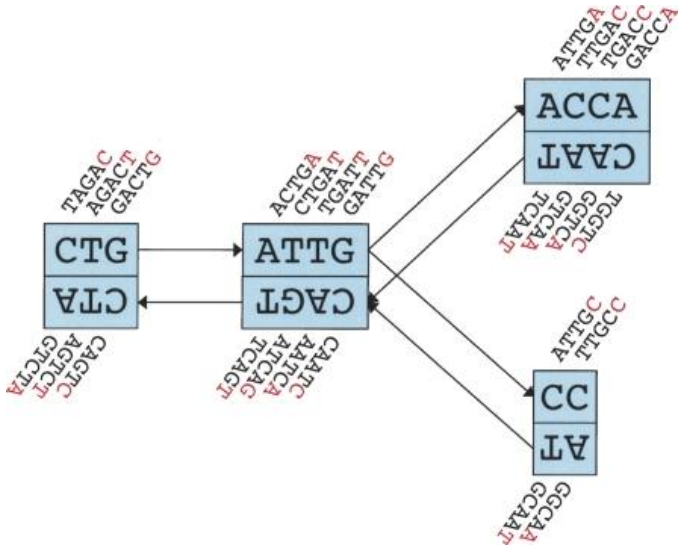

FIGURA 13 – REPRESENTAÇÃO ESQUEMÁTICA DO GRAFO DE BRUIJN. CADA NÓ, REPRESENTADO POR UM ÚNICO RETÂNGULO, É RESULTANTE DE UMA SÉRIE DE SOBREPOSIÇÕES DE K-MERS (NESTE CASO, K=5). O ÚLTIMO NUCLEOTÍDEO DE CADA K-MER (EM VERMELHO). AS SEQUÊNCIAS DE NÚCLEOTÍDEOS DESSE FINAL (LETRAS DENTRO DO RETÂNGULO) É A SEQUÊNCIA DO NÓ. O NÓ DUPLO, LIGADO DIRETAMENTE AOS NÓS, ABAIXO E ACIMA, REPRESENTA A SÉRIE COMPLEMENTAR INVERSA DOS K-MERS.

FONTE: ZERBINO e BIRNEY, 2008.

O Velvet tem em sua estrutura básica dois módulos de execução: Velveth e Velvetg. O Velveth tem como função principal a criação de uma tabela de índices (*hash*) a partir de um conjunto de seqüências de leituras, computando sobreposições entre k-mers, e gerar no final do processo dois arquivos principais: um arquivo de

seqüências indexadas (*sequence*) e um arquivo contendo a representação das sobreposições entre os k-mers chamado de mapa de vias (*Roadmaps*).

O módulo Velvetg trabalha na construção propriamente dita do genoma através da manipulação do grafo de Bruijn, correção de erros e resolução de repetições, gerando um arquivo de contigs ou scaffolds (regiões contínuas com seqüências de bases determinadas) resultantes da sobreposição das seqüências, a descrição plena do grafo de Bruijn produzido (*LastGraph*), a descrição das ações executadas (*Log*), estatísticas da montagem (Stats.txt) e opcionalmente, um arquivo das seqüências (*UnusedReads.fa*).

Os parâmetros de montagem a serem definidos pelo usuário são:

- -shortMatePaired: Indica o tipo de biblioteca de entrada (Mate-pair)
- -cov_cuttof; Remoção de nós/arcos de baixa cobertura
- -insp_length: Distância esperada entre pares
- -min_contig_lgth: Tamanho mínimo para consenso
- -exp_cov: Estimativa da cobertura para regiões únicas.

Para execução das montagens deste trabalho foram utilizados os parâmetros mostrados nas TABELAS 9, 10, 11, 12.

TABELA 10 - PARÂMETROS UTILIZADOS NO MÓDULO VELVETH DO MONTADOR VELVET

Parâmetros	**Valor**
Tipo de Biblioteca	Curto-Pareado
Tamanho do K-mer	21

TABELA 11 - PARÂMETROS UTILIZADOS NO MÓDULO VELVETG DO MONTADOR VELVET

Parâmetros	**Valor**
Distância entre os pares	336pb
Cobertura esperada	32x
Cobertura mínima	3
Tamanho mínimo de contig	1000pb
Criação de scaffolds	sim

TABELA 12 - PARÂMETROS UTILIZADOS NO MÓDULO VELVETH DO MONTADOR VELVET PARA UMA MONTAGEM ALTERNATIVA.

Parâmetros	**Valor**
Tipo de Biblioteca	Curto-Pareado
Tamanho do K-mer	21

TABELA 13 - PARÂMETROS UTILIZADOS NO MÓDULO VELVETG DO MONTADOR VELVET PARA UMA MONTAGEM ALTERNATIVA.

Parâmetros	**Valor**
Distância entre os pares	336pb
Cobertura esperada	32x
Cobertura mínima	3
Tamanho mínimo de contig	sem tamanho mínimo
Criação de scaffolds	não

3.3.4 De novo Assembly

A sequência de programas para a montagem (FIGURA 21) *de novo* com os dados Solid é denominada *pipeline de novo* Assembly e contém um conjunto de ferramentas criadas e adaptadas que permitem a reconstrução de genomas através de leituras gerada pelo seqüenciador Solid da empresa Applied Biosystems (SOLID™4). Este *pipeline* é utilizado por causa do alto número de seqüências produzidas pelo seqüenciador Solid dos valores de qualidade para sequência codificadas em cores (*colorspace*), bem como do esquema de codificação em 2-bases para correta montagem das leituras. Este *pipeline* também utiliza grafos de Bruijn implementados pelo Velvet (APPLIED BIOSYSTEMS).

O programa SAET (SOLID™ Accuracy Enhancement Tool), que é uma ferramenta de correção de erros das leituras SOLID, também foi utilizado. O programa ASID (Assembly assistant for SOLID™) foi utilizado para correção de falhas ou lacunas de montagem (*gaps*).

Para avaliar as montagens obtidas foram utilizados os seguintes parâmetros:

- Contig N50: 50% das bases montadas estão representados em um contig deste determinado tamanho e maiores (quanto maior o N50 melhor a montagem)

- Scaffold N50: A soma das bases em scaffolds de tamanhos iguais ou maiores representa 50% do tamanho total dos scaffolds.
- Tamanho médio do contig (valores mais altos são melhores)
- Tamanho do maior contig (valores mais altos são melhores)
- Número de gaps (falhas): Regiões do genoma não cobertos pela montagem (menos gaps indicam melhor montagem)
- Número de contigs: > 100 bases (contigs de tamanho grande, de preferência)

A montagem de um genoma é um processo multi-paramétrico complexo que depende do tamanho do genoma, a sua complexidade, a cobertura, o tempo de leitura e a precisão das seqüências obtidas para a montagem.

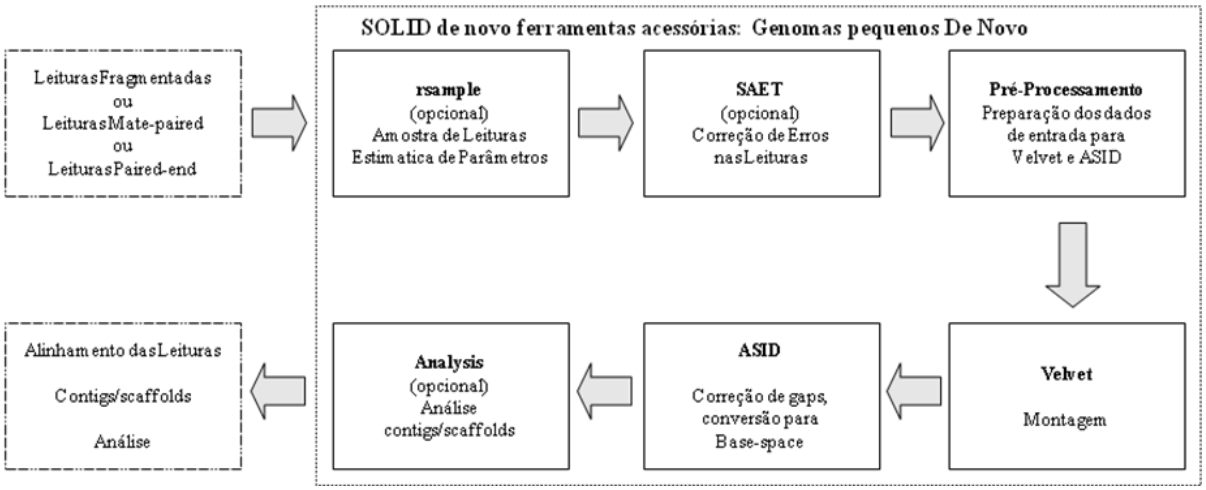

FIGURA 14 – FLUXO DE EXECUÇÃO DE MONTAGEM DO PIPELINE DE NOVO ASSEMBLY.

FONTE: Adaptação (De Novo Assembly Protocol)

O processo de execução mostrado na FIGURA 15 descreve as etapas que envolvem o processo de reconstrução do genoma. O início se dá em uma etapa opcional por parte do usuário, onde parâmetros são estimados para correção de erros, avançando para a pré-montagem propriamente dita pelo aplicativo SAET (SOLID™ Accuracy Enhancement) que faz a preparação e conversão dos arquivos do formato atual (csfasta) para um formato padrão compatível para a montagem (double encoded). Na etapa seguinte o Velvet cria os grafos de Bruijn, para que na próxima etapa o ASID (Assembly assistant for SOLID™) proceda com a correção

das falhas de montagem (gaps) ocorridas durante o processo, e converta do formato adaptado (*double encoded*) para o formato de seqüência de nucleotídeos (*base-space*), gerando os resultados finais (APPLIED BIOSYSTEMS). Para esta montagem foram adotados os parâmetros listados na TABELA 14.

TABELA 14 - PARÂMETROS UTILIZADOS PARA O PROCESSO DE EXECUÇÃO NO PIPELINE DE NOVO ASSEMBLY.

Parâmetros	**Valor**
Tamanho de K-mer	21
Tamanho estimado do genoma	7.5Mpb*
Distância entre os pares	1500pb
Cobertura esperada	32x
Variação de distância entre os pares	500pb
Tamanho da leitura	35pb
Máximo de cobertura	800x
Número de processadores	16

*****Tamanho estimado do genoma foi baseado no organismo de referência.

3.3.5 jContigSort

O jContigSort é uma ferramenta desenvolvida em plataforma Java para ordenar os contigs com base no genoma de referência. O algoritmo cria um conjunto de todas as sementes possíveis de sequência de DNA do genoma de referência. As sementes e sua posição no genoma são armazenadas em uma tabela hash (espalhamento) (CORMEN *et al.* 2002). As semente criadas a partir dos contigs são pesquisadas na tabela hash (espalhamento) e as posições das sementes (que possuem a mesma seqüência) no genoma de referência são associadas ao contig. O conjunto das posições é ordenado e obtém-se a mediana das posições e este valor é associado ao contig. Posteriormente, os contigs serão ordenados pelos valores das medianas obtida.

A orientação para cada contig é obtida pela comparação do número de partidas em cada vertente do genoma de referência (GUIZELINI *et al.* 2011).

3.3.6 BLAST

O BLAST (Basic Local Alignment Search Tool) é uma ferramenta de comparação rápida de seqüências. Seu algoritmo simples pode ser aplicado de

várias formas e em uma variedade de contextos como buscas em banco de dados de proteínas, busca e identificação de genes e análise de regiões com semelhança em longas seqüências. O BLAST tornou-se uma das ferramentas de comparação de busca e comparação de seqüências existente mais utilizada (ALTSCHUL ET AL. 1990).

3.3.7 Mummer

A comparação de genomas tem sido um método utilizado para se compreender funções gênicas e a evolução do genoma desde o inicio do processo de seqüenciamento. O Mummer é uma ferramenta criada para comparações de genomas, tanto eucarióticos como procarióticos. Através dos gráficos de alinhamentos gerados torna-se possível observar a distância evolutiva em múltiplos genomas (KURTZ *et al*. 2004).

3.3.8 MATLAB

O MATLAB® pode ser definido como uma linguagem de alto nível de computação técnica em um ambiente interativo de algoritmos, visualização de dados e computação numérica. Ele pode ser usado em uma ampla gama de aplicações, entre elas processamento de sinal e imagem e biologia computacional com uma caixa de ferramentas com diversas funções para fins especiais. O MATLAB pode ser usado para resolver problemas técnicos de computação mais rápidos e ágeis do que com linguagens tradicionais de programação, como C, C++ e Fortran, podendo ainda ser integrado com outras linguagens e aplicações (MathWorks)

Neste trabalho o MATLAB foi utilizado como ferramenta básica para criar scripts necessários para a solução de pequenos problemas, para visualização do alinhamento de pequenas regiões e cálculo e estatística básica da montagem.

3.3.9 RAST

O RAST (Rapid Annotations Using Subsystems Technology) é um serviço totalmente automatizado para anotação rápida do genoma de bactérias e arqueas. Ele identifica a codificação de proteínas, rRNA, atribui funções a genes, prevê que

subsistemas estão representados no genoma, e usa essas informações para reconstruir a rede metabólica. Facilita para o usuário visualizar e comparar o genoma anotado mantido no ambiente SEED. O serviço tem sido utilizado pela comunidade científica e já foi usado para anotação de mais de 350 genomas distintos (AZIZ *et al*., 2008).

3.4 ESTRATÉGIA DE MONTAGEM

Como estratégia principal de montagem foram utilizadas duas técnicas distintas de seqüenciamento para que então os resultados pudessem se complementar (FIGURA 16).

Após a obtenção dos dados brutos de seqüenciamento, análises iniciais foram efetuadas para a determinação de possíveis erros provenientes do seqüenciamento e assim determinar possível filtragem dos dados e definição de melhores parâmetros de montagem. Para o conjunto de dados Ilumina foram repetidamente testados parâmetros com o objetivo de obter-se uma montagem satisfatória, onde o objetivo foi alcançar o menor número de scaffolds possível e o número de bases totais o mais próximo do tamanho do genoma do organismo de referência.

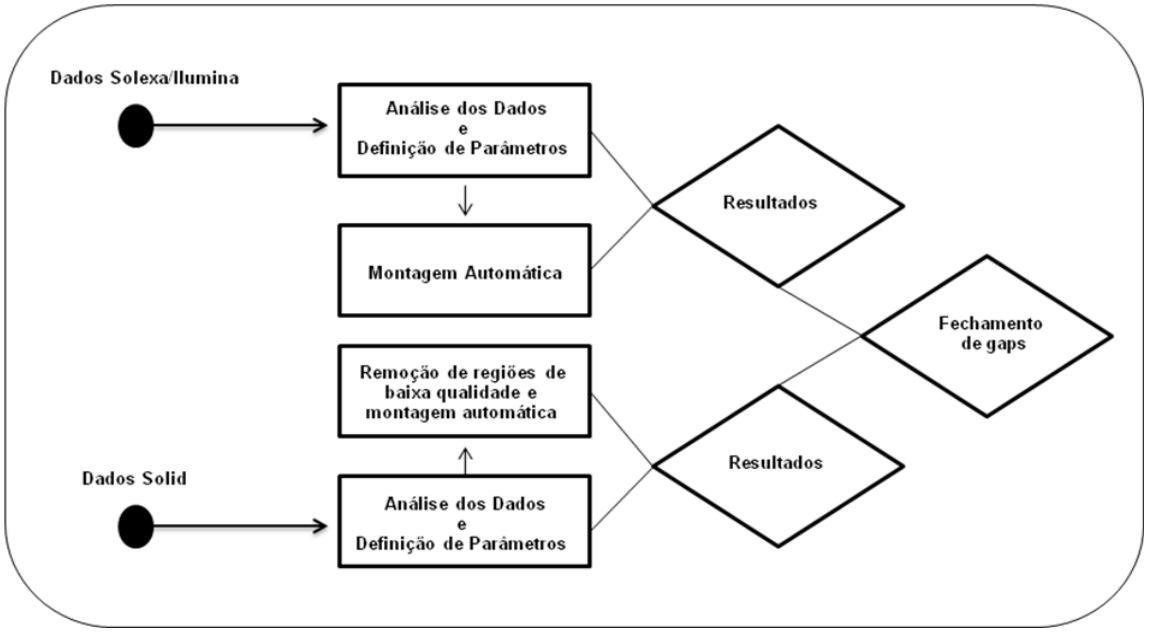

FIGURA 15 – FLUXOGRAMA GERAL DO PROCESSO DE MONTAGEM

FONTE: O Autor (2012)

Após análise dos dados Solid, optou-se pela remoção de região das seqüências consideradas de baixa qualidade (Item 3.3.1). Uma limitação técnica de hardware impediu que repetidos testes fossem feitos para obtenção de melhores resultados, portanto poucas tentativas foram feitas, as duas montagens foram mescladas com o objetivo de correção de falhas (gaps).

3.4.1 Remoção de regiões de baixa qualidade

O processo de filtragem é uma etapa fundamental na reconstrução de um genoma (montagem), uma vez que reduz a freqüência de alinhamentos incorretos que são normalmente causados por erros nas leituras e também pelo tamanho das leituras. A partir das análises dos dados pelos aplicativos FastQC (Solexa) e Quality Assessment (Solid) foi possível determinar alguns problemas no conjunto dos dados. Os dados Solid apresentaram bases repetidas na extremidade direita (5') em grande parte das leituras do conjunto dos dados e regiões onde bases não foram identificadas pelo seqüenciador (substituídas por pontos).

Por isso foram retiradas em 15 pb da extremidade 5', transformando as leituras de 50 pb em 35 pb. Para isso foi utilizado script em linguagem Perl levando-se em consideração o tipo e o formato do arquivo. Não foi necessário realizar a filtragem dos dados Ilumina ou remoção de regiões de baixa qualidade.

3.4.2 Estratégia de correção de falhas de montagem

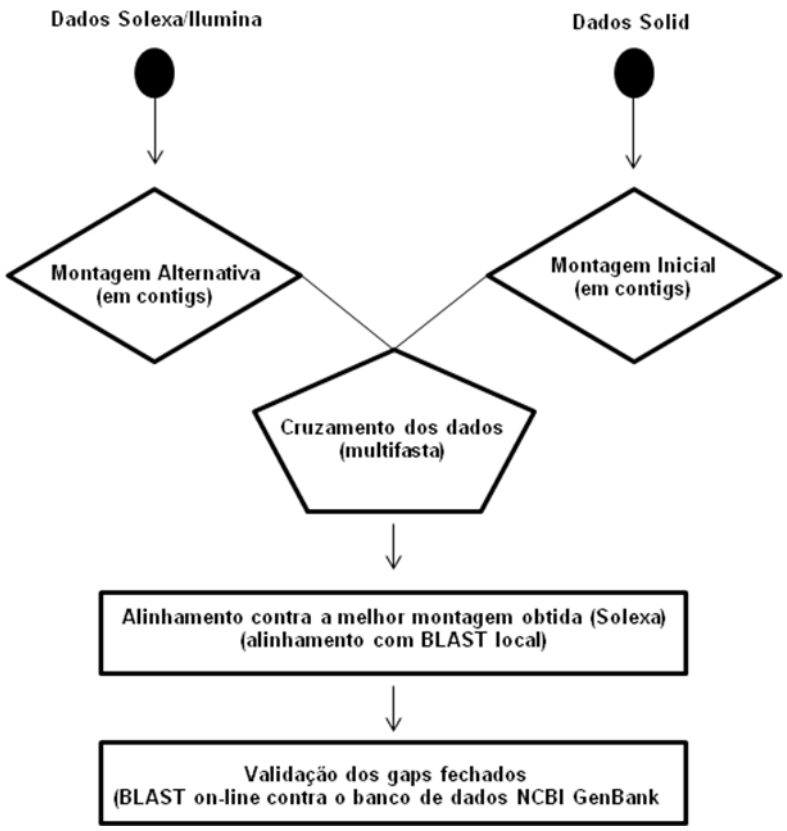

FIGURA 16 – FLUXOGRAMA GERAL DO PROCESSO DE FECHAMENTO DE FALHAS DE MONTAGEM.

FONTE: O Autor (2012)

Para o processo de correção de falhas de montagem, adotou-se como estratégia principal a mesclagem das montagens de ambas as tecnologias de seqüenciamento e alinhamento contra a montagem mais confiável e de melhor resultado estatístico (FIGURA 24). Para os dados Solexa foram criadas duas montagens com critérios de avaliação diferentes, a primeira onde o objetivo era obter a melhor montagem considerando os critérios de avaliação já mencionados (Item 3.3). Esta montagem foi chamada de montagem 1. Foi também realizada uma montagem 2, onde os critérios de avaliação seria a quantidade máxima de leituras incluídas para manter a máxima quantidade de informação na montagem. Essa segunda montagem com os dados Illumina obtida foi utilizada juntamente com a

montagem Solid para compor um arquivo multifasta que foi então utilizado para alinhar com o programa BLAST com os contigs/scaffolds da melhor montagem, em termos de número de falhas e tamanho de contigs/scaffolds, obtidas com os dados Ilumina.

Para esta tarefa foi utilizado o software de alinhamento BLASTN localmente, onde o arquivo multifasta (montagem 2 contigs Solid) foi transformado em banco de dados e o conjunto de contigs da montagem 1 foi utilizada como seqüência query. A FIGURA 18 mostra um resultado deste processo onde foi encontrado um contig no banco de dados que completa uma falha da montagem 1. O fechamento da falha foi feito manualmente, após análise visual, em editor de texto (gedit). Após o fechamento da falha, a nova sequência foi verificada utilizando BLAST contra o banco de dados NR do GenBank (FIGURA 19). As validações dos fechamentos das falhas foram feitas de forma amostral aleatória.

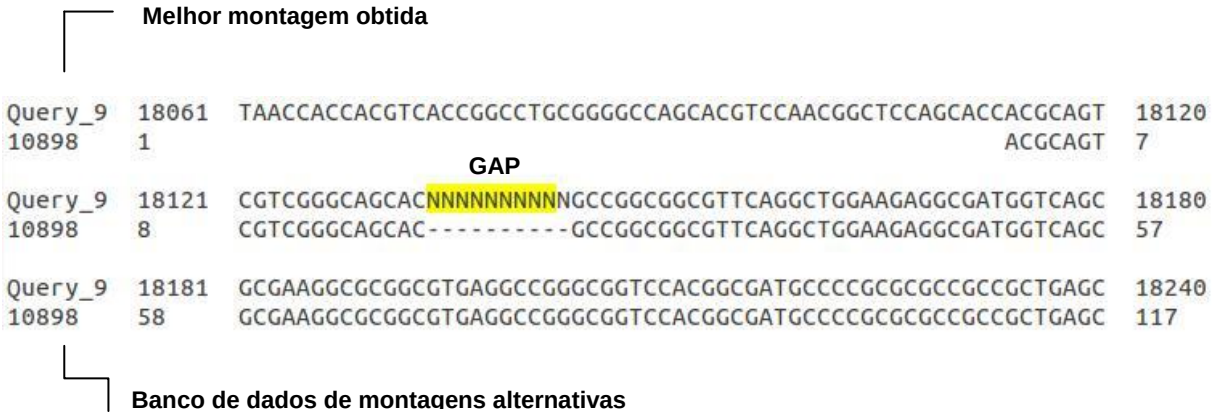

FIGURA 17 - REPRESENTAÇÃO DO ALINHAMENTO E FECHAMENTO DE UMA LACUNA OU FALHA. MONTAGENS ALTERNATIVAS FORAM CRIADAS COM O OBJETIVO DE CORRIGIR AS FALHAS TRAZIDAS PELO MONTADOR VELVET (GAPS), ESSAS MONTAGENS FORAM ALINHADAS CONTRA A MELHOR MONTAGEM OBTIDA (QUARY), OS NÚMEROS DA QUERY SÃO AS POSIÇÕES DE ALINHAMENTO NA MELHOR MONTAGEM OBTIDA E O NÚMERO ABAIXO FOI O CONTIG ALINHADO DENTRO DAS MONTAGENS ALTERNATIVAS (P.EX. CONTIG 10898).

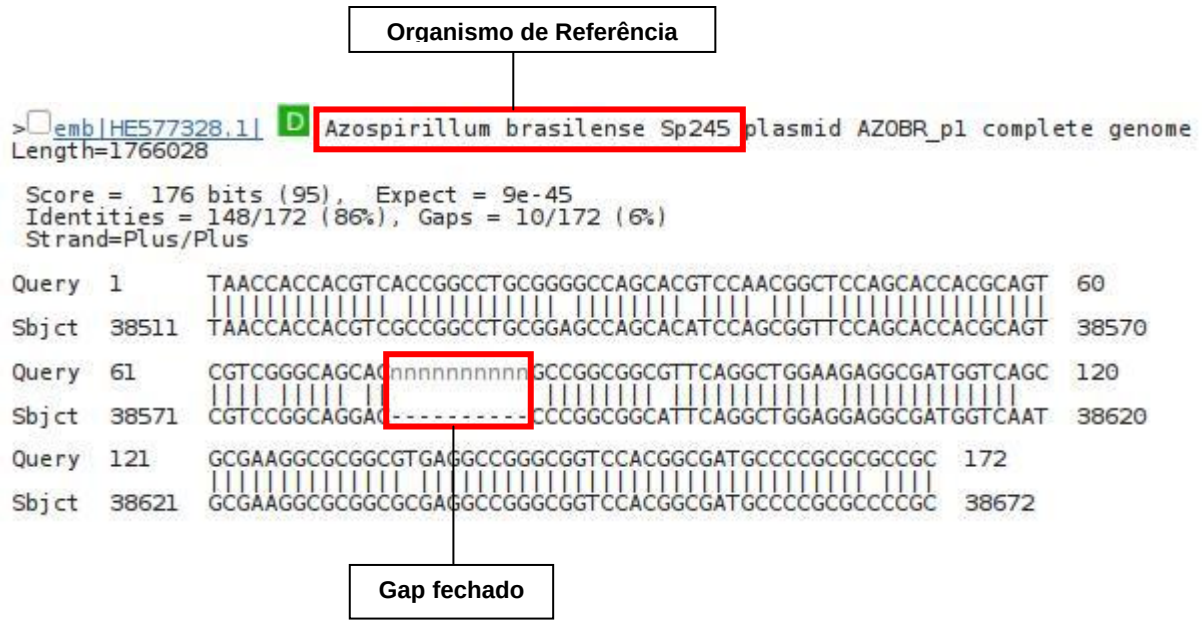

FIGURA 18 - REPRESENTAÇÃO DO RESULTADO DE VALIDAÇÃO DO FECHAMENTO DO GAP. A REGIÃO ONDE O ALINHAMENTO COM AS MONTAGENS ALTERNATIVAS PERMITIRAM CORRIGIR AS FALHAS DE MONTAGEM, FORAM VALIDADAS ATRAVÉS DO BLAST CONTRA O GENBANK.

4 RESULTADOS E DISCUSSÃO

4.1 ANÁLISE DE DADOS BRUTOS

4.1.1 Dados Ilumina

O gráfico de qualidade das leituras illumina apresenta altos valores de qualidade em praticamente todas as bases, mantendo sempre o valor mediano, o valor inter-quartil e a qualidade média na região considerada de muito boa qualidade (verde), mas mostrando uma queda acentuada no nível nos valores de qualidade nas bases finais das leituras (phred 37-36). Mesmo assim, permanece dentro da região considerada de muito boa qualidade, sem cair para as regiões chamadas de qualidade razoável (laranja) ou de má qualidade (vermelho) (FIGURA 20)

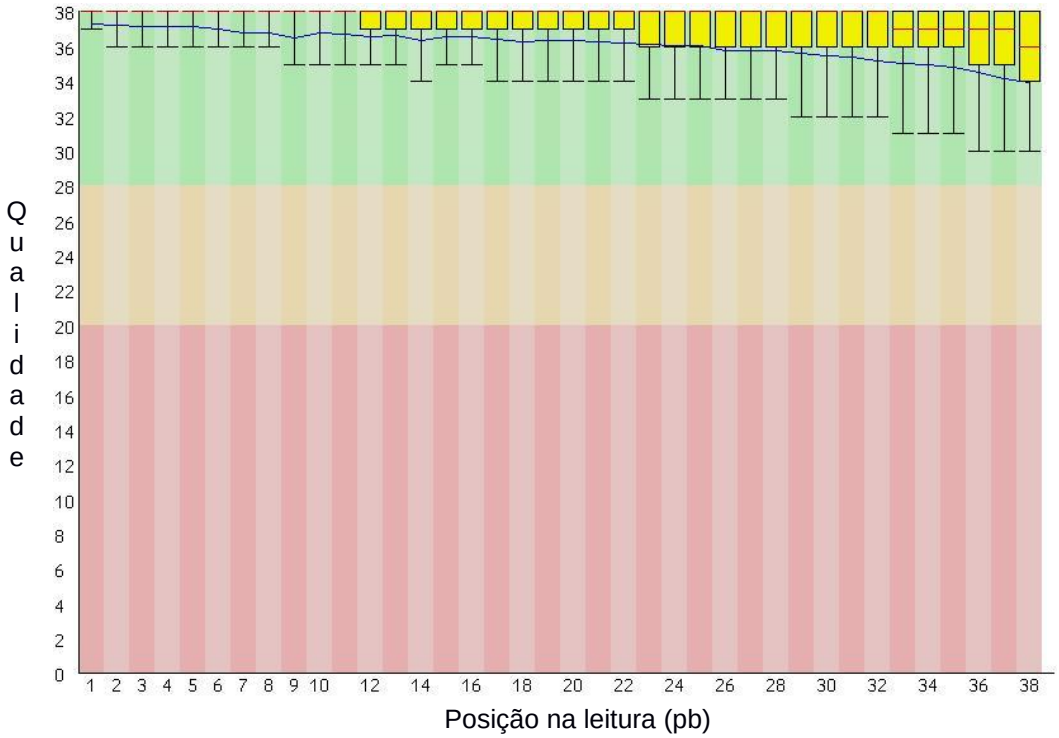

FIGURA 19 – REPRESENTAÇÃO GRÁFICA DA QUALIDADE MÉDIA POR BASE DOS DADOS ILLUMINA. VERDE: BOA QUALIDADE; LARANJA: QUALIDADE RAZOÁVEL; VERMELHO: QUALIDADE RUÍM

A maior parte das seqüências derivadas do sequenciador Illumina tinha escore de qualidade médio de 37 ou superior. Quase a totalidade das seqüências tinha qualidade média entre 34 e 38 (FIGURA 21)

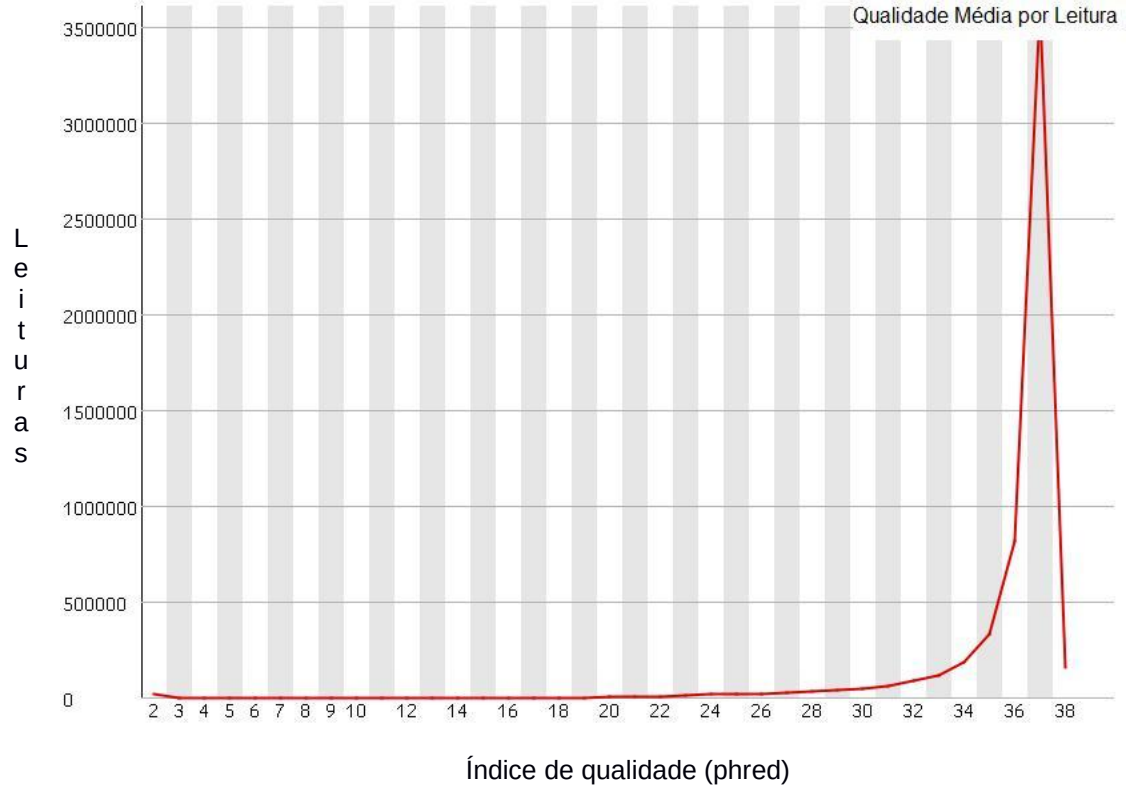

FIGURA 20 - REPRESENTAÇÃO GRÁFICA DA DISTRIBUIÇÃO DA QUALIDADE MÉDIA SOBRE TODAS AS SEQUÊNCIAS.

Quando a distribuição de qualidade base-a-base das seqüências (FIGURA 22), foi avaliado nota-se que há uma quantidade maior de bases G (32%) e C (32% em relação a A (18%) e T (18%) no conjunto analisado. A distribuição das bases nas leituras é homogenia, mas há uma leve desproporção no inicio das seqüências (10 primeiras bases), mostrando uma super-representação de todas as bases nesta região.

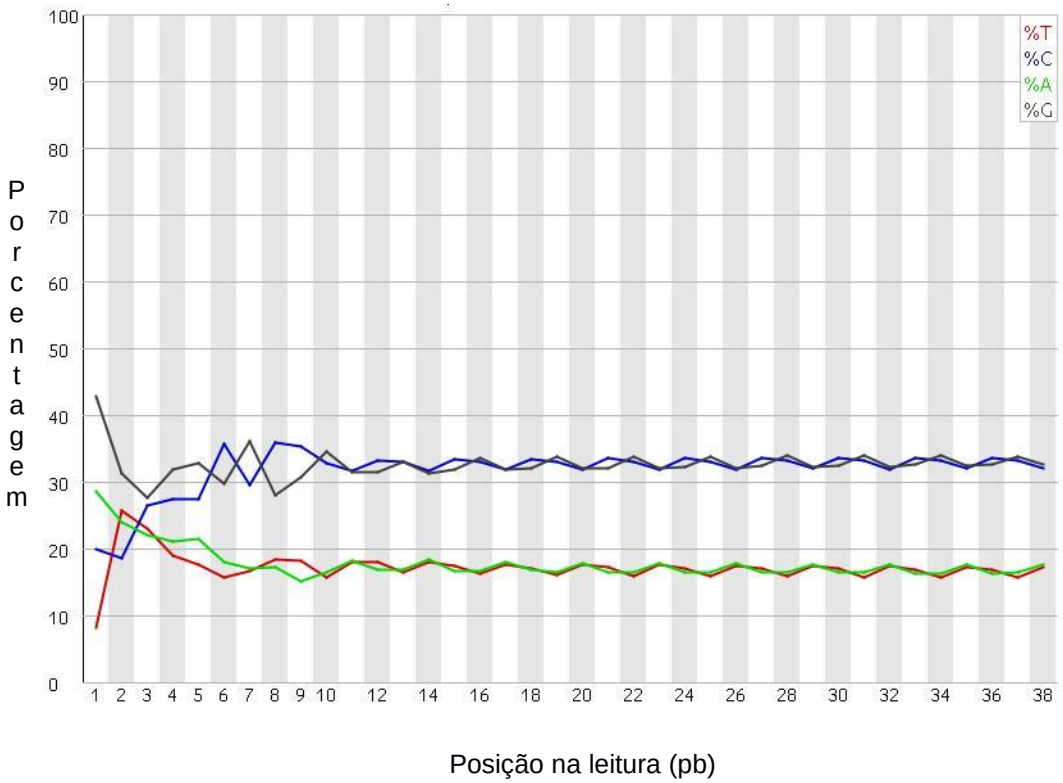

FIGURA 21 - REPRESENTAÇÃO GRÁFICA DA DISTRIBUIÇÃO POR BASE DO CONJUNTO DE SEQUÊNCIAS ILLUMINA.

Na avaliação do conteúdo de GC sobre todas as bases (FIGURA 23), e na distribuição de GC sobre todas as seqüências, a mesma desproporção no inicio das seqüências mostrada na FIGURA 22 se repete. A FIGURA 23 mostra que as bases G e C em todas as seqüências representam aproximadamente 67%. Isto mostra uma distribuição normal esperada, onde o pico central que corresponde o conteúdo total de GC do genoma está em conformidade com a distribuição teórica do modelo apresentado pelo módulo de análise (FIGURA 24). Para esse conjunto de dados não foi necessário qualquer processo de filtragem baseado em poda das seqüências por baixa qualidade.

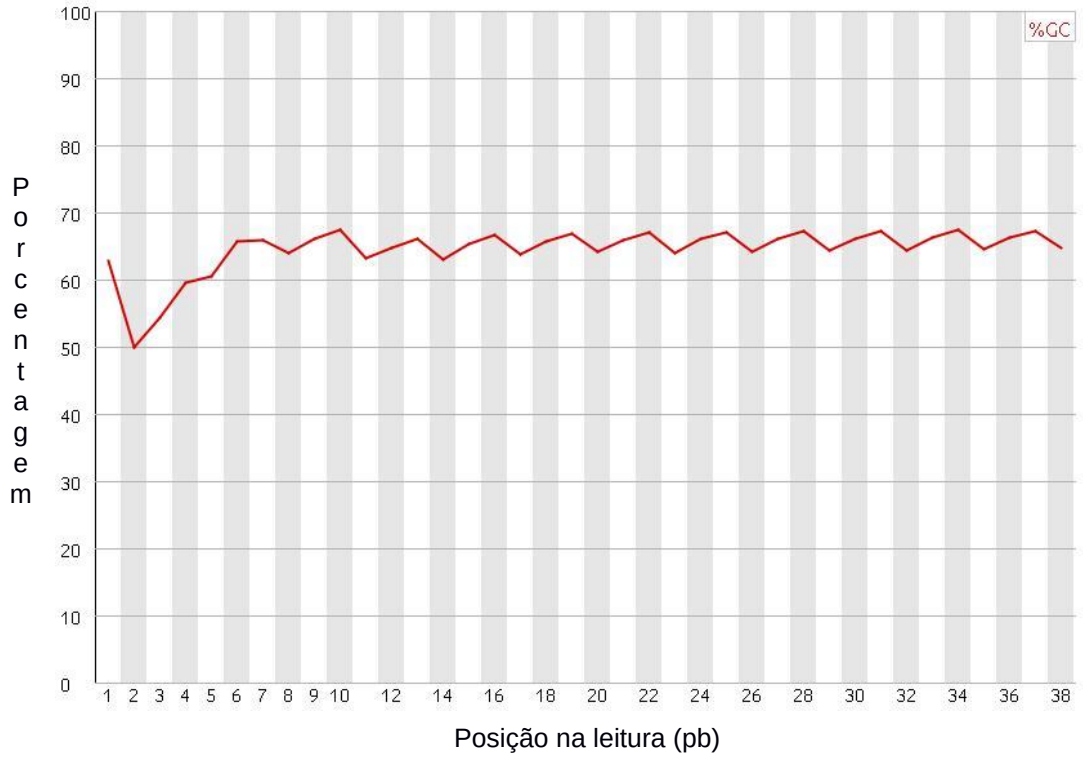

FIGURA 22 – REPRESENTAÇÃO GRÁFICA DO CONTEÚDO DE GC SOBRE TODAS AS BASES.

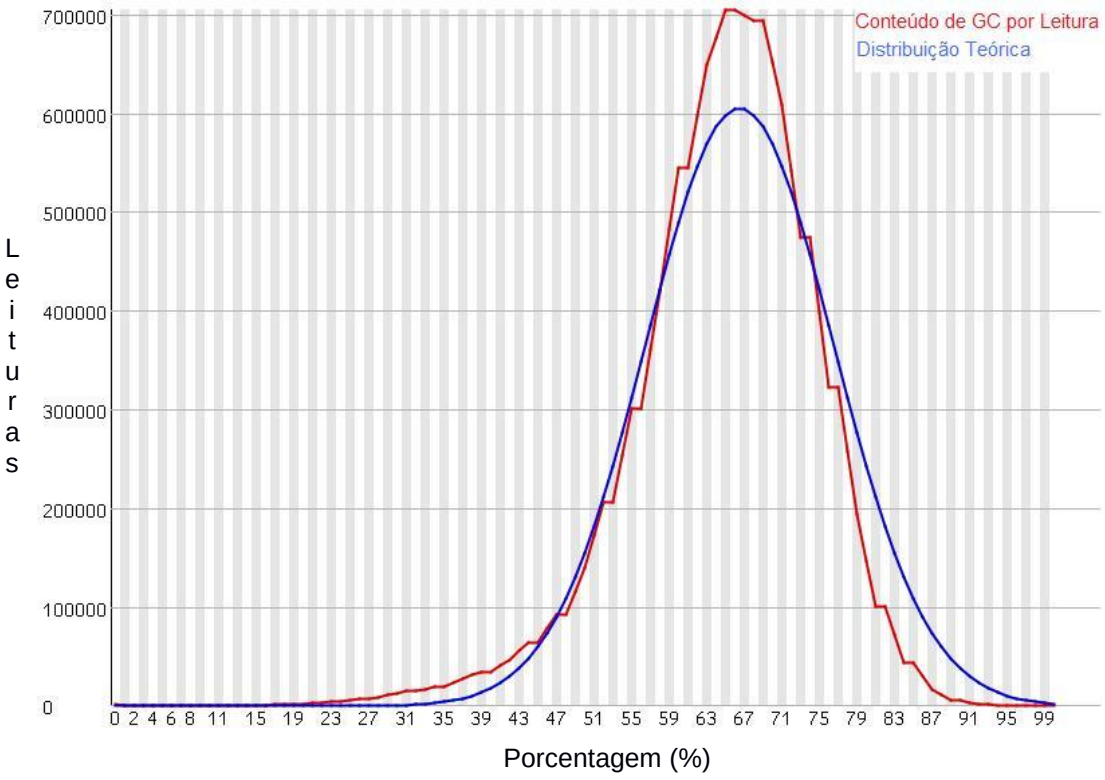

FIGURA 23 – REPRESENTAÇÃO GRÁFICA DO CONTEÚDO DE GC SOBRE TODAS AS SEQÜÊNCIAS.

4.1.2 Dados SOLID

As seqüências SOLID são obtidas a partir das duas extremidades com primers diferentes, chamados F3 e R3. A representação gráfica das seqüências F3 e R3 do conjunto de dados SOLID (FIGURA 25 e 26) mostraram uma distribuição dos valores de qualidade por base com variações substanciais em todo o conjunto. Nas seqüências F3 (FIGURA 25), a qualidade média foi 14, obtendo um pico máximo de qualidade de phred 17 no inicio das seqüências e uma queda substancial de qualidade no final das seqüencias.

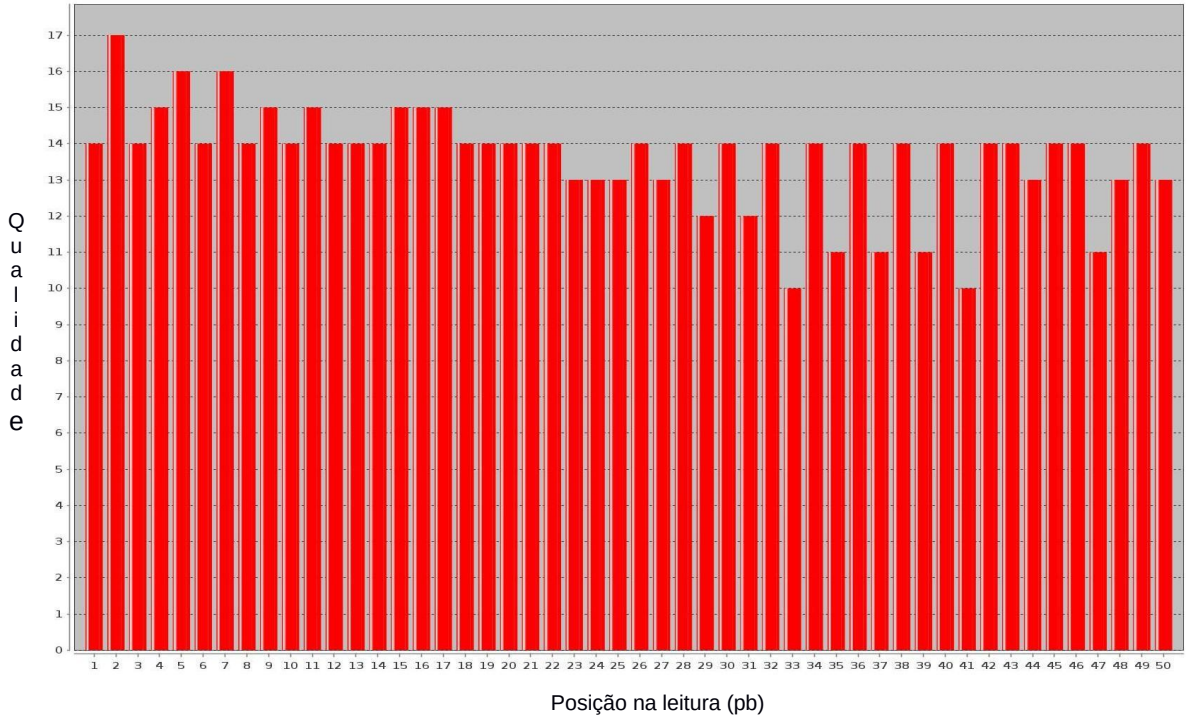

FIGURA 24 – REPRESENTAÇÃO GRÁFICA DA DISTRIBUIÇÃO DE QUALIDADE BASE-A-BASE NAS SEQUÊNCIAS OBTIDAS COM PRIMER F3.

Quanto às seqüências R3 (FIGURA 26), a qualidade varia por toda a extensão das seqüências, onde ocorrem variações com picos de phred 18 (no início das seqüências) e uma queda da posição 21 em diante com picos de phred 17 e 18 no final das seqüências. Estas variações surgem um provável problema ocorrido no processo de seqüenciamento.

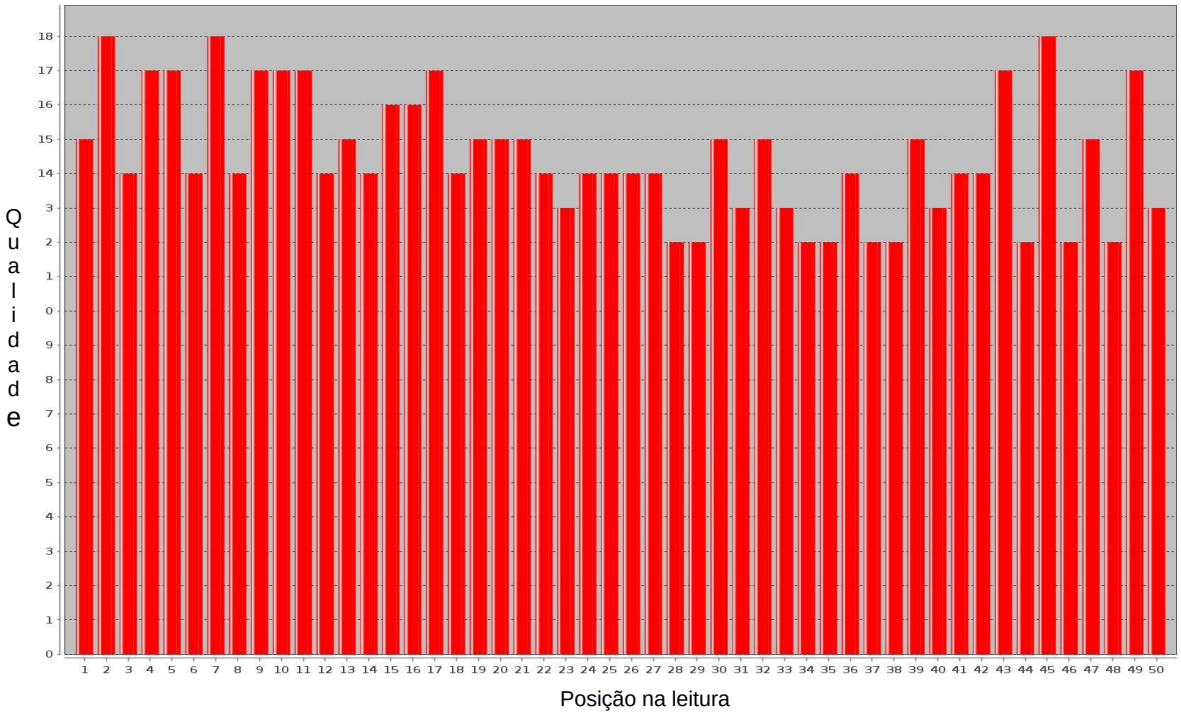

FIGURA 25 - REPRESENTAÇÃO DA DISTRIBUIÇÃO DE QUALIDADE BASE-A-BASE NAS SEQÜÊNCIAS OBTIDAS COM PRIMER R3.

Estes problemas podem ser visualizados quando observamos os dados brutos em formato original (csfasta) ou convertidos para bases (nucleotídeos) (FIGURA 27 e 28). Os pontos mostrados na FIGURA 27, que apresentam perdas de sinal nesta posição, são menos freqüentes nas seqüências R3 do que nas F3. Além disso, as sequência também apresentam uma grande quantidade de bases repetitivas, normalmente na extremidade (5') direita das leituras (FIGURA 28).

```
>427_6_1717_R3
G00231203010100102.033301021.11000200000001.00.1.00
>427_6_1722_R3
G01010332030332000.002013003.00030000300000.00.0.00
>427_6_1743_R3
G00000003300000020.000000000.00000000000000.00.0.00
```

FIGURA 26 – EXEMPLOS DE SEQUÊNCIAS COM O PRIMER R3 NO FORMATO ORIGINAL (CSFASTA).

```
>428_9_1380_R3
GTCCTTTTGGCGCCACGCATTATTAGGAACTGGGGACCCCCAAAAAAAAA
>428_9_1388_R3
GCGTGTAGCATCCGGTGATGCAACAGCTCTCGGAGACTTTTGAGGGGAAA
>428_9_1427_R3
ATGAACTGCTTGAAGACCGGCTTCTCCTCCTCCGACTCCTCGAAAACTTT
>428_9_1459_R3
ACCCTGACCGGGCTGTCTCACACCAACGCCCAATCTAAAATAAAAAAGGG
>428_9_1523_R3
CGGGGTCCTTGATGTGGAAGCGATATTCGCGCCCCGCTTTTCTTTTTCCC
>428_9_1607_R3
AGGCGCGGCCCTTTCTTTCCAGGGGGAAAAAAAAAAAAAAACCCCCCCCC
```

FIGURA 27 – EXEMPLOS DE SEQUÊNCIAS COM O PRIMER R3 NO FORMATO ORIGINAL (CSFASTA) CONVERTIDOS PARA NÚCLEOTÍDEOS (FASTA).

Nas seqüências F3, os pontos mencionados anteriormente são mais freqüentes, o que inviabiliza o uso de parte do conjunto de dados. Contudo, pela grande cobertura que traz o seqüenciamento da tecnologia SOLID, espera-se que os problemas de pontos existentes em parte dos dados, como mostrado na FIGURA 29, sejam sanados no processo de construção do genoma.

```
>427_9_45_F3
T33..233..333..331..332..020..203...21...22....2...
>427_9_66_F3
T23..133..232..210..030..230..220...22...20....2...
>427_9_94_F3
T20..333..333..300..110..200..300...21...21....2...
>427_9_160_F3
T21..110..301..332..210..223..311...22...22....2...
```

FIGURA 28 – EXEMPLOS DE SEQUÊNCIAS COM O PRIMER F3 NO FORMATO ORIGINAL (CSFASTA).

Sendo assim, optou-se por remover as regiões com problemas identificados pelas análises. Foram retiradas 15 bases da extremidade direita (5') de todas as leituras (FIGURA 28), deixando cada seqüência com 35 pb.

4.2 RESULTADOS DE MONTAGEM

O processo da primeira montagem automática utilizando as seqüências Illumina, onde o objetivo foi obter o genoma (6.7 Mpb) em um menor número de contigs/scaffolds. Os parâmetros estatísticos desta montagem 1 estão mostrados na TABELA 15.

TABELA 15 - ESTATÍSTICA DE MONTAGEM AUTOMÁTICA PARA OS DADOS ILLUMINA.

Característica do conjunto de dados	**Scaffolds**
Tamanho total do genoma (pb)	6.941.807 pb
Número de scaffolds	444
Bases indeterminadas (pb)	~123.648 pb
Número de gaps	5.531
Scaffold N50	26.262
Maior scaffold (pb)	176.802 pb
Quantidade de leituras usadas	94,7%

O tamanho do genoma nesta montagem foi um pouco maior do que o estimado (6.7 Mpb) por Martin-Didonet *et al.* (2000) para a espécie *Azospirillum brasilense* FP2 (Item 1.1). O *scaffold* N50 tem aproximadamente 26 Kbp, ou seja, cerca de 3.8 Mpb do genoma estão contidos nos scaffolds maiores que 26 Kpb. Os parâmetros estatísticos da montagem com os dados de origem SOLID, estão na TABELA 16.

TABELA 16 - ESTATÍSTICA DE MONTAGEM AUTOMÁTICA PARA OS DADOS SOLID

Característica do conjunto de dados	**Scaffolds**	**Contigs**
Tamanho total do genoma	7.1 Mpb	7.1 Mpb
Número de *scaffolds*/contigs	4717	10024
Número de gaps	~5545 pb	0
Scaffold/contig N50	53947 pb	1453 pb
Maior *scaffold*/contig	199110 pb	22025 pb
Quantidade de leituras usadas	48,5%	48,5%

O tamanho do genoma estimado foi 7,1 Kpb, valor muito próximo do estimado anteriormente (TABELA 15). Embora o número total de *scaffolds* tenha sido muito alto (~4700), o scaffold N50 foi aproximadamente 50 Kpb e o maior scaffold tinha cerca de 199 Kpb.

4.2.1 Montagem Illumina alternativa

Como estratégia para a correção das falhas foi realizada uma montagem alternativa, onde o critério adotado foi utilizar não só as leituras já utilizadas na montagem anterior (TABELA 15), bem como, as leituras que não entraram na montagem anterior pelo rigor dos parâmetros adotados. Sendo assim obtivemos uma segunda montagem cujo resultado está mostrado na TABELA 17.

TABELA 17 - ESTATÍSTICA DA MONTAGEM 2

Característica do conjunto de dados	**contigs**
Tamanho total do genoma (pb)	7,1 Mpb
Número de contigs	7482
Contig N50	3421
Maior scaffold (pb)	40028 pb
Quantidade de leituras usadas	96,5%

O tamanho do genoma estimado nesta montagem alternativa, chamada de montagem 2, foi de 7.1 Mpb (TABELA 17), e houve um acréscimo de leituras que não tinham sido incluídas na montagem 1.

4.4 FECHAMENTO DE FALHAS DE MONTAGEM

A estratégia adotada nos permitiu visualizar cada região onde a falha ocorreu, validar a estratégia e fechar parte das falhas. Para isso, um arquivo multifasta foi criado, contendo os resultados apresentados na montagem SOLID em contigs (10024) e o resultado da montagem 2 dos dados Illumina (7482), gerados para esse fim. Este arquivo contém 17.506 contigs, que foram alinhados contra os contigs da montagem 1. A TABELA 18 mostra os resultados que obtivemos após a estratégia que adotamos para a correção das falhas.

TABELA 18 – ESTATÍSTICA DO FECHAMENTO DE FALHAS DE MONTAGEM

Falhas de Montagem (gaps)	**Corrigidos**	**Restantes**
5531	2060	3470
	Scaffolds	**Contigs**
Antes das correções	433	10
Depois das correções	386	57

Antes do processo de correção das falhas a montagem 1 apresentava 433 seqüências onde haviam regiões de falhas (regiões contínuas de bases separadas por N's), denominadas *scaffolds* e 10 regiões contínuas de bases sem falhas (contigs), totalizando 433 *scaffolds* e 10 contigs. Após o processo de correções 47 dos *scaffolds* foram transformados em contigs, totalizando 57 contigs e restando ainda a ser corrigidos 386 scaffolds com falhas.

Algumas regiões necessitaram ser analisadas de forma mais pontual, uma vez que o processo de validação revelou discordância. Nas regiões onde os alinhamentos foram perfeitos e inequívocos (FIGURA 30 e 31), foi adotada a substituição direta como processo de correção (retirar os N's ou substituí-los pelas bases alinhadas com as montagens alternativas). Nas regiões onde houve divergência (FIGURA 32 e 33) foi realizada uma análise visual da região, com busca e comparação com seqüências homólogas no banco de dados GenBank. Em alguns casos não havia falhas nas regiões alinhadas (FIGURA 34).

***Azospirillum brasilense* FP2**

```
Query_24  115    GAAGTCGGCGCGTCCGACCACCACCTCGTCCAGCCGCATCTCCCACAGGGCGGTGGTGCC  174
7012      2338   GAAGTCGGCGCGTCCGACCACCACCTCGTCCAGCCGCATCTCCCACAGGGCGGTGGTGCC  2279

Query_24  175    GGGATCGACCAACGTGGGCNNNNNNNNNGCAGCCCGGCGGGGGTGGGCACCAGCCACTT  234
7012      2278   GGGATCGACCAACGTGGGCGCGGCGCCGCGCAGCCCGGCGGGGGTGGGCACCAGCCACTT  2201
```

FIGURA 29 – ALINHAMENTO DOS *SCAFFOLDS* CONTRA AS MONTAGENS ALTERNATIVAS (BLAST LOCAL). QUERY: MONTAGEM 1; ARQUIVO MULTIFASTA CRIADO PARA O ALINHAMENTO: 1 À 10.024 (SEQUÊNCIAS SOLID); 10.024 À 17.506 (SEQUÊNCIAS DA MONTAGEM ALTERNATIVA ILUMINA).

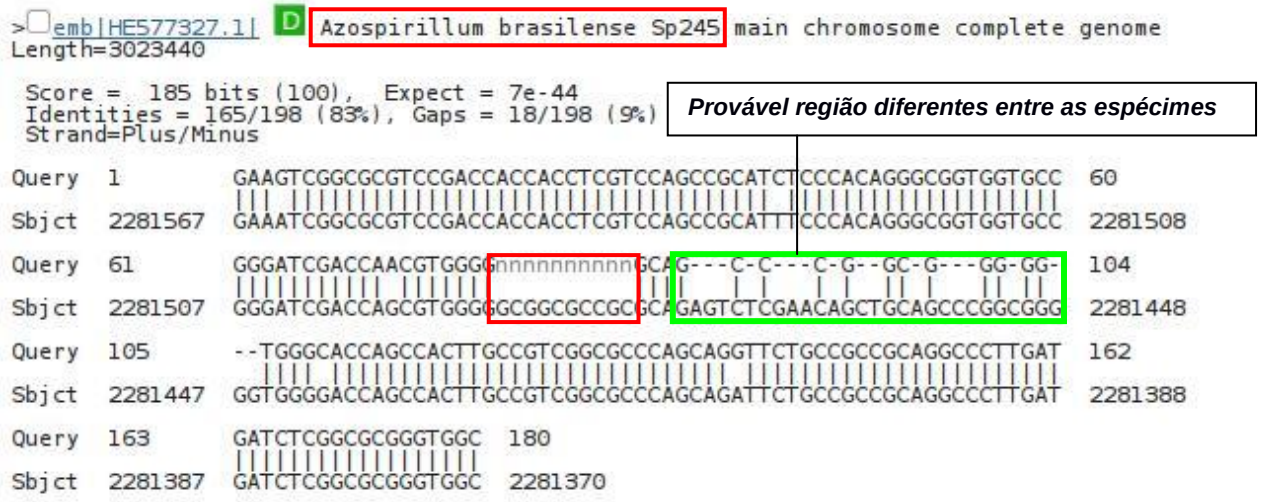

FIGURA 30 - COMPARAÇÃO DA REGIÃO COM FALHA COM O BANCO DE DADOS GENBANK, MOSTRANDO A SEQUÊNCIA DO GENOMA DE REFERÊNCIA.

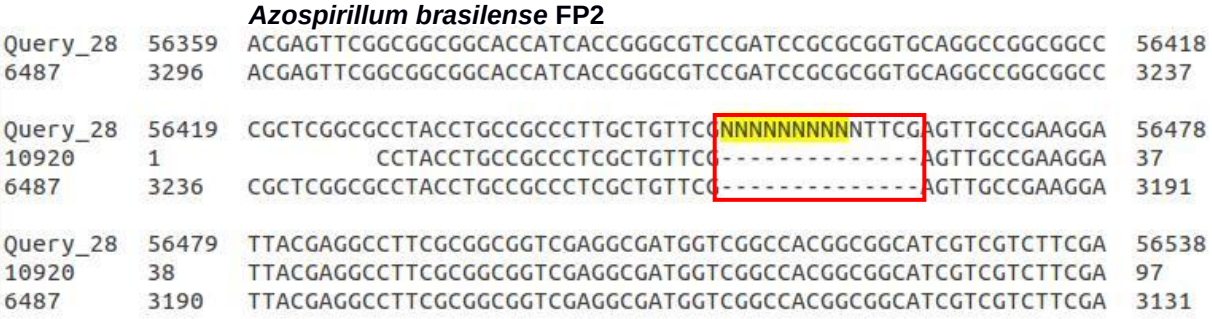

FIGURA 31 – ALINHAMENTO DOS SCAFFOLDS CONTRA AS MONTAGENS ALTERNATIVAS (BLAST LOCAL). QUERY: MONTAGEM 1; ARQUIVO MULTIFASTA CRIADO PARA O ALINHAMENTO: 1 À 10.024 (SEQUÊNCIAS SOLID); 10.024 À 17.506 (SEQUÊNCIAS DA MONTAGEM ALTERNATIVA ILLUMINA).

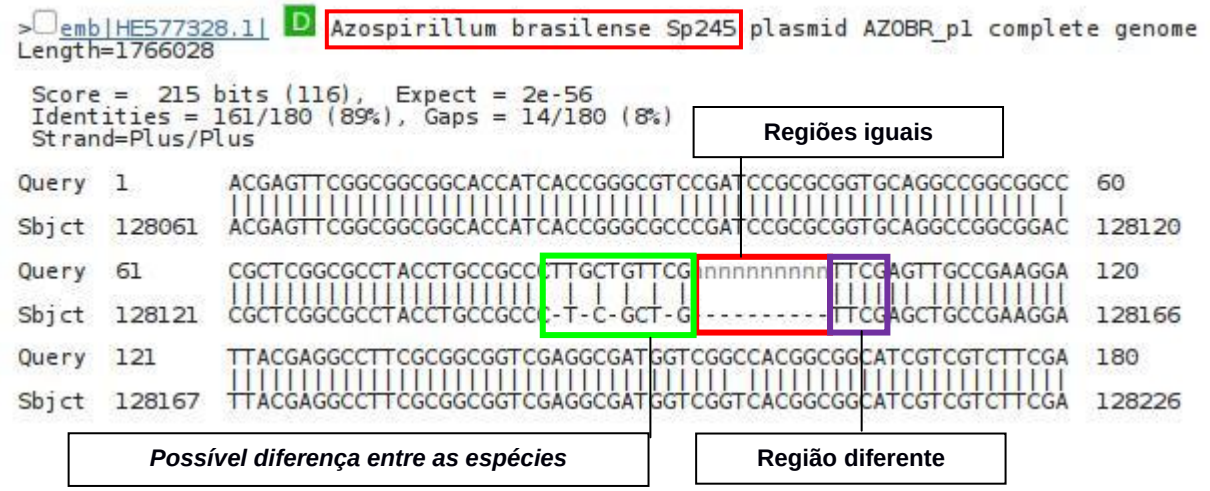

FIGURA 32 – FALHAS DE MONTAGEM COM REGIÕES IGUAIS E DIFERENTES ENTRE A ANOTAÇÃO E O ALINHAMENTO ENTRE AS MONTAGENS.

FIGURA 33 – REGIÕES DE FALHAS DE MONTAGEM SEM ALINHAMENTO.

4.5 PROCESSO DE ORDENAÇÃO

A ordenação foi baseada no organismo de referência *Azospirillum brasilense* Sp245. Para a ordenação foram usadas sementes de 21 pb. O alinhamento foi visualizado em um gráfico de alinhamento gerado pelo Mummer (FIGURA 35).

O gráfico apresenta no eixo x a sequência do organismo de referência (*Azospirillum brasilense* Sp245) e no eixo y o genoma da bactéria *Azospirillum brasilense* FP2. A linha diagonal central indica a existência de alta identidade entre

duas seqüências em posições equivalentes no genoma, revelando portanto um alto grau de co-linearidade dos genes nos genomas das duas espécies.

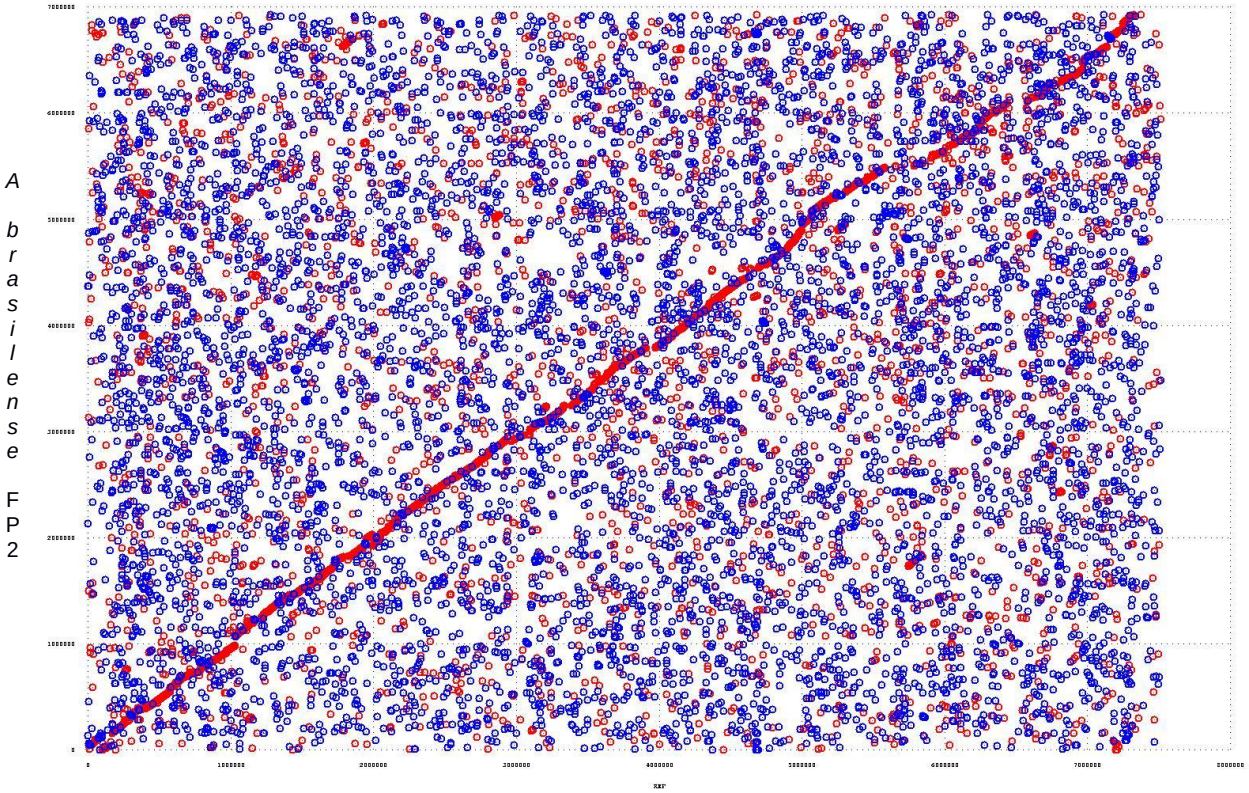

FIGURA 34 – ALINHAMENTO DO GENOMA COM O ORGANISMO DE REFERÊNCIA, APÓS O PROCESSO DE ORDENAÇÃO, UTILIZADO O CONJUNTO TOTAL DA MONTAGEM (444 SCAFFOLDS).

4.6 MONTAGEM FINAL

A última versão da montagem do genoma de *A. brasilense* FP2 representa uma melhora substancial em relação ao resultado inicial (TABELA 15), principalmente no que tange a correção das falhas de montagem (TABELA 19).

TABELA 19 – ESTATÍSTICA FINAL DE MONTAGEM

Característica do conjunto de dados	Scaffolds
Tamanho total do genoma (pb)	6.941.807 para 6.931.925 pb
Número de scaffolds/contigs	386 / 57
Bases indeterminadas restantes (N)	~103.630 pb
Número de bases indeterminadas corrigidas (N)	20.018
Número de gaps restantes	3470
Scaffold N50	26.262
Maior scaffold (pb)	176.802 pb
Total de bases A	1.089.764 nt
Total de bases T	1.095.035 nt
Total de bases G	2.322.465 nt
Total de bases C	2.321.031 nt
Percentual de GC	68 %

Com a estratégia adotada para o fechamento das falhas de montagem foi possível corrigir 2060 gaps, diminuindo o número de falhas no genoma para 3470 gaps. O tamanho apresentou uma pequena diminuição (0.15%) em relação ao tamanho inicial (TABELA 15), devido a remoção dos Ns (20.018) e adição de novas bases. O genoma final apresenta 68% de GC, valor muito próximo do organismo de referência (68.4%).

4.6.1 Alinhamento do *A. brasilense* FP2 com o genoma de referência

Os dados brutos Solexa e Solid foram comparados com a seqüência genômica de *A. brasilense* Sp245 com a utilização da ferramenta MosaikAligner seguindo valores padrões sugerido pelo protocolo de utilização. O resultado mostrou que 13% das leituras dos dados Solid alinharam ao genoma de referência enquanto que 78% das seqüências Illumina foram alinhadas (TABELAS 20 e 21).

4.6.2 Alinhamento aos genes 16S e 23S rRNA da espécie de referência

Os genes 16S e 23S rRNA apresentaram alinhamento total dentro da sequência obtida da espécie de *A. brasilense* FP2. No entanto, apresentaram alinhamento em apenas uma região dentro da sequência total o que mostrou discrepância com os resultados obtidos na espécie de referência e os resultados de anotação.

TABELA 20 – MAPEAMENTO DOS DADOS BRUTOS SOLID COM O GENOMA DE REFERÊNCIA.

Estatística de Alinhamento	
Falha de Hash	45237692 (40,1 %)
Filtrado fora	59828735 (53,0 %
Único	7402877 (6,6 %)
Não único	312721 (0,3 %)
Total	112782.028
Total alinhado	**7715598 (6,8 %)**
Estatisticas de Alinhamento (leituras)	
Não Alinhados	49068779 (87,0 %)
Deixado de fora	6928872 (12,3 %)
Mates únicos	359575 (0,6 %)
Total de leituras	56391014
Total de leituras alinhadas	**7322235 (13,0%)**
Alinhamento de mate em pb	**378064302**
Candidatos alinhados	**246544**

Os resultados de alinhamentos da TABELA 20 mostram que o mapeador identificou um total 112.782.028 leituras em pares (mate-paired), sendo que desse total 6,6 % foram identificados como regiões únicas de alinhamento e 3,3 % regiões não únicas. O mapeador deixou de fora por falha de hash 40,1 % e pelo parâmetro de filtro 53 %, resultando em um total de seqüências em pares alinhado ao genoma de referência de 6,8 % do total identificado. Quando o analisado foi leitura sem levar em conta os pares, foram identificadas 56.391.014 leituras, sendo que 87% do total não foi alinhados e 12,3 % deixado de fora, resultando em um alinhamento de 13 % do total de leituras identificadas.

TABELA 21 - MAPEAMENTO DOS DADOS BRUTOS ILLUMINA COM O GENOMA DE REFERÊNCIA

Estatística de Alinhamento (mates)	
Seqüências deixadas de fora	1927880 (33.5 %
Único	3613170 (62.7 %)
Não único	220456 (3.8 %)
Total	5.761.506
Total alinhado	**3833626 (66.5 %)**
Estatística de Alinhamento (leituras)	
Não Alinhados	633330 (22.0 %)
Deixado de fora	668102 (23.2 %)
Pares de mates únicos	1444554 (50.2 %)
Um mate não único	80528 (2.8 %)
Pares de mates não únicos	54680 (1.9 %)
Total de leituras	2884194
Total de leituras alinhadas	**2250864 (78.0%)**
Alinhamento de mate em pb	**145649602**
Candidatos alinhados	**136763**

Para os dados Illumina mostrados na TABELA 21, o mapeador identificou um total 5.761.506 leituras em pares (leituras pareadas), sendo que desse total 62.7 % foram identificados como regiões únicas de alinhamento e 3.8 % regiões não únicas. O mapeador deixou de fora pelo parâmetro de filtro 33.5 % do total, resultando em um total de seqüências em pares alinhadas ao genoma de referência de 66.5 % do total identificado. Quando o analisado foi leitura sem levar em conta os pares, foram identificadas 2.884.194 leituras, sendo que 22% do total não foram alinhados e 23.2 % deixado de fora, resultando em um alinhamento de 78.0 % do total de leituras identificadas.

4.7 ANÁLISE PARCIAL DA ANOTAÇÃO

A anotação realizada pelo programa RAST identificou 6119 regiões codificadoras (genes). Destas, 2327 regiões foram alocadas a subsistemas e 3792 regiões não foram alocadas a nenhum subsistema. Os subsistemas podem ser definidos como vias metabólicas (por exemplo, glicólise), ou seja, conjunto de proteínas funcionais que realizam um processo biológico ou compõem uma estrutura complexa.

TABELA 23 – RESULTADOS DE ANOTAÇÃO DA SEQUÊNCIA GENÔMICA PARCIAL DE *A. brasilense FP2.*

Resumo de Análise

Genoma	*Azospirillum brasilense* FP2
Domínio	Bactéria
Taxonomia	Bactéria; Proteobactéria; Alphaproteobactéria; Rhodospirilales; Rhodospirillaceae; *Azospirilum brasilense*;
Espécies próximas	*Azospirillum sp*. B510 *Rhodospirillum centenum* SW *Magnetospirillum magneticum* AMB1 *Magnetospirillum gryphiswaldense* MSR1 *Rhodospirillum rubrum* ATCC 11170
Tamanho, pb	6,931,925
Subsistemas	2327
Não pertencentes a subsistemas	3792
Sequências codificadoras	6119
RNAs	60
Possíveis genes faltando	200

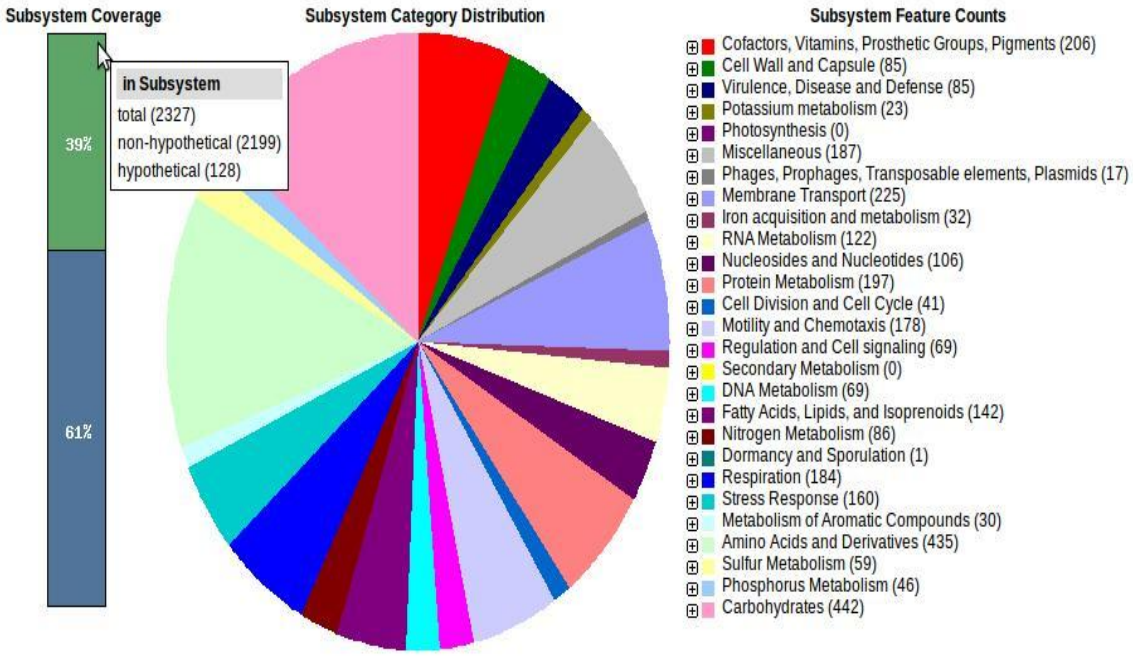

FIGURA 35 – ESTATÍSTICA DE DISTRIBUIÇÃO DE CATEGORIAS DE SUBSISTEMAS DOS RESULTADOS GERADOS PELO RAST ON-LINE.

O gráfico apresentado na FIGURA 35 apresenta um conjunto de genes classificados funcionalmente (subsistemas), ou seja, regiões que codificam proteínas que possivelmente estão ligadas a um processo biológico específico (via metabólica). Os dados mostram que dos 6119 genes encontrados, 39% estão classificados em grupos de famílias de proteínas ligadas a possíveis processos biológicos relacionados de vias metabólicas, desse total de 2327 pertencentes a subsistemas, 94,4% foram consideradas como proteínas não hipotéticas e 5,5 % como proteínas hipotéticas. Do total de regiões codificadoras 61 % estão classificadas em regiões não pertencentes a subsistemas metabólicos, dos quais 44,3 % são de proteínas não hipotéticas e 55,6 % de proteínas hipotéticas.

Os 2327 genes identificados como subsistemas parecem codificar proteínas diretamente ligadas a processos bioquímicos de transporte de membrana (sistema de secreção de proteínas – TypeII, transportadores ABC), metabolismo do nitrogênio (fixação de nitrogênio, amonificação de nitrato e nitrito e assimilação de amônia) entre outros (FIGURA 36).

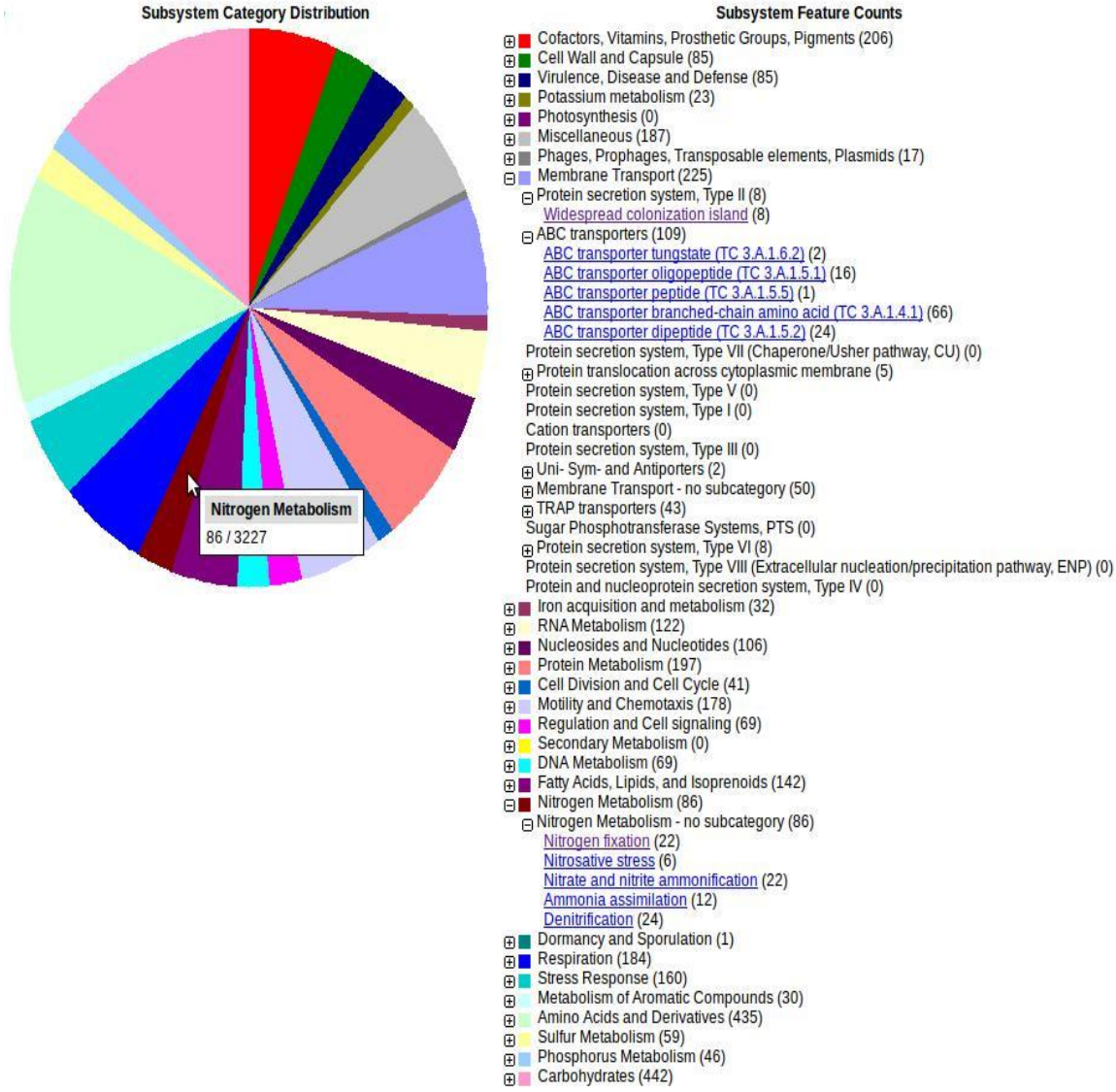

FIGURA 36 - RESULTADO DA DISTRIBUIÇÃO DE SUBSISTEMAS GERADOS PELO RAST AUTOMÁTICAMENTE.

A TABELA 24 mostra alguns dos genes envolvidos com a fixação biológica de nitrogênio do subsistema metabolismo de nitrogênio. Neste subsistema também são encontrados genes envolvidos com estresse nitrosativo, amonificação de nitrato e nitrito, assimilação de amônia e desnitrificação.

TABELA 24 – GENES ENVOLVIDOS COM FIXAÇÃO BIOLÓGICA DE NITROGÊNIO DE *A. brasilense* FP2

Categoria	Subsistema	Função
Metabolismo do nitrogênio	Fixação de Nitrogênio	Nitrogenase (molibdênio-ferro) regulador específico, NifA
Metabolismo do nitrogênio	Fixação de Nitrogênio	Nitrogenase Fe-Mo, cofator para estrutura e montagem, proteína NifE
Metabolismo do nitrogênio	Fixação de Nitrogênio	Proteína associada a nitrogenase proteína NifO
Metabolismo do nitrogênio	Fixação de Nitrogênio	Nitrogenase, montagem de FeMo cofator proteína NifB
Metabolismo do nitrogênio	Fixação de Nitrogênio	Nitrogenase, cadeia Alfa (molibdênio-ferro)
Metabolismo do nitrogênio	Fixação de Nitrogênio	Proteína NifX
Metabolismo do nitrogênio	Fixação de Nitrogênio	Cisteína desulfurase, NifS
Metabolismo do nitrogênio	Fixação de Nitrogênio	Proteína protetora da Nitrogenase NifW
Metabolismo do nitrogênio	Fixação de Nitrogênio	Ferrodoxina associada a nitrogenase
Metabolismo do nitrogênio	Fixação de Nitrogênio	Homocitrato sintase
Metabolismo do nitrogênio	Fixação de Nitrogênio	Proteína NifQ
Metabolismo do nitrogênio	Fixação de Nitrogênio	Proteína NifZ
Metabolismo do nitrogênio	Fixação de Nitrogênio	Proteína NifT
Metabolismo do nitrogênio	Fixação de Nitrogênio	Nitrogenase cadeia beta, cadeia beta da proteína molibdênio-ferro
Metabolismo do nitrogênio	Fixação de Nitrogênio	Proteína NifU

5 CONCLUSÕES

1. A seqüência parcial do genoma da bactéria fixadora de nitrogênio *Azospirillum brasilense* estirpe FP2 foi obtida utilizando duas plataformas de seqüenciamento de nova geração, Ilumina e Solid, que produzem leituras curtas de 38 e 50pb, respectivamente.

2. O genoma de *Azospirillum brasilense* FP2 tem aproximadamente 7 Mpb.

3. A utilização das duas plataformas de seqüenciamento de nova geração Ilumina e Solid, permitiu o fechamento de falhas.

4. A anotação automática da sequência genômica parcial revelou que o *Azospirillum brasilense* FP2 codifica cerca de 6119 genes.

REFERÊNCIAS

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., & LIPMAN, D. J. **Basic Local Alignment Search Tool**. *Journal of Molecular Biology, v. 215*, p. 403-410, 1990.

APPLIED BIOSYSTEMS. **De novo assembly protocool**. Life Technologies Corporation, 2010.

AZIZ, R. K.; BARTELS, D.; BEST, A. A.; DEJONGH, M.; DISZ, T.; EDWARDS, R. A.; FORMSMA, K.; GERDES, S.; GLASS, E. M.; KUBAL, M.; MEYER, F.; OLSEN, G. J.; OLTON, R.; OSTERMAN, A. L.; OVERBEEK, R. A.; MCNEIL, L. K.; PAARMANN, D.; PACZIAN, T.; PARRELLO, B.; PUSCH, G. D.; REICH, C.; STEVENS, R.; VASSIEVA, O.; VONSTEIN, V.; WILKE, A.; ZAGNITKO, O. **The RAST server: Rapid Annotation using subsystems technology.** BMC Genomics, v. 8, p. 75, 2008.

BABRAHAM BIOINFORMATICS Disponível em: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/ - Acesso em: 06/01/12

BEN DEKHIL, S., CAHILL, M., STACKEBRANDT, E.; SLY, L. I. **Transfer of Conglomeromonas largomobilis subsp. Largomobilis to the genus Azospirillum as Azospirillum largimobile comb. nov., and elevation of Conglomeromonas largomobilis subsp. Parooensis sp. nov.** Syst. Appl. Microbiol., v. 20, p. 72-77, 1997.

BALDANI, J. I.; CARUSO, L.; BALDANI, V. L. D.; GOI, S. R.; DÖBEREINER, J. **Recent advances in BNF with non-legume plants**. Soil Biology & Biochemistry, Oxford, v. 29, n. 5/6, p. 911-922, 1997.

BARAK, R., NUR, I., OKON, Y., HENIS, Y. **Aerotactic response of *Azospirillum brasilense*.** J. Bacteriol. v. 152, p. 643-649, 1982.

CARVALHO, M. C. C. G.; SILVA, D. C. G. **Seqüenciamento de DNA de nova geração e suas aplicações na genômica de plantas**. Ciência Rural, Santa Maria, v. 40, n. 3, p. 735-744, 2010.

CHAN, E. Y. **Advances in sequencing technology**. Mutation Research, v. 573, p. 13-40, 2005.

CORNER, T.H.; LEISERSON, C.E.; RIVEST, R.L.; STEIN, C. **Algoritmos: teoria e prática**. Elsevier, 11ª reimpressão, Rio de Janeiro, 2002.

CHOU, H. H.; HOLMES, M. H. **DNA sequence quality trimming and vector removal**. Bioinformatics. V. 17, p. 1093-1104, 2001.

DIDONET, A. D.; MARTIN-DIDONET, C. C. G.; GOMES, G. F. **Avaliação de linhagens de arroz de terras altas inoculadas com *Azospirillum lipoferum* Sp59b e *A. brasilense* Sp245**. Comunicado técnico EMBRAPA, p. 1678-961X, 2003.

DÖBEREINER, J.; PEDROSA, O. P. **Nitrogen-fixing bacteria in non-leguminous crop plants. Berlin:** Springer-Verlag, 1987. 155 p. 1987.

DÖBEREINER, J., BALDANI, V.L.D. and REIS, V.M. **Endophytic occurrence of diazotrophic bacteria in non-leguminous crops**, 1995. In: **Azospirillum VI and Related Microorganisms** (FENDRIK, I., DEL GALLO, M.; VANDERLEYDEN, J.; ZAMAROCZY, M., (Eds.), Springer, Berlin, p. 3-14, 1995.

DURFEE, T.; NELSON, R.; BALDWIN, S.; PLUNKETT, G.; BURLAND, V.; MAU, B.; PETROSINO, J. F.; QIN, X. MUZNY, D. M.; AYELE, M.; GIBBS, R. A.; CYÖRGO, B.; PÓSFAI, G.; WEINSTOCK, G. M.; BLATTNER, F. R. **The complete genome sequence of Escherichia coli DH10B: Insights into the biology of a laboratory workhorse.** Journal of bacteriology, v. 190, n. 7, p. 2597-2606, 2008.

ECKERT, B.; WEBER, O. B.; KIRCHHOF, G.; HALBRITTER, A.; STOFFELS, M.; HARTMANN, A. ***Azospirillum doebereinerae* sp. nov., a nitrogen-fixing bacterium associated with the C_4-grass *Miscanthus***. Int J Syst Evol Microbiol, v. 51, p. 17-26, 2001.

EWING, B.; HILLIER, L.; WENDL, M.; GREEN P. **Base-Calling of automated sequencer traces phred**. I. Using Assessment. Genome Research, v. 8, p. 175-185, 1998.

FEDURCO, M.; ROMIEU, A.; WILLIANS, S.; LAWRENCE, I.; TURCATTI, G. **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.** Nucleic Acids Research, v. 17, n. 1, p. 69-73, 2006.

GUIZELINI, D.; PEDROSA, F.O.; MARCHAUKOSKI, J.N.; STEFFENS, M.B.R.; SOUZA, E.M.; TIBÃES, J.H.; SOUZA, V. AND RAITTZ; R. T. **JContigSort: the computer application for contigs ordering. In: 7th International Conference of the Brazilian Association for bioinformatics and Computacional Biology (AB3C) and 3rd International Conference of the IberoAmerican Society for Bioinformatics (SolBio),** Florianópolis, abstract book, 2011.

HALL, P.G. and KRIEG, N.R. **Application of the indirect immunoperoxidase stain technique to the flagella of *Azospirillum brasilense***. Appl. Environ. Microbiol. v. 47, p. 433-435, 1984.

KENEKO, T.; MINAMISAWA, K.; ISAWA, T.; NAKATSUKASA, H.; MITSUI, H.; KAWAHARADA, Y.; NAKAMURA, Y.; WATANABE, A.; KAWASHIMA, K.; ONO, A.; SHIMIZU, Y.; TAKAHASHI, C.; MINAMI, C.; FUJISHIRO, T.; KOHARA, M.; KATOH, M.; NAKAZAKI, N.; NAKAYAMA, S.; YAMADA, M.; TABATA, S. **Complete genomic structure of the cultivated rice endophyte *Azospirillum* sp. B510**. DNA Research, v. 17, p. 37-50, 2010.

KENNEDY, I.R.; CHOUDHURY, A. T. M. A.; MIHA'LY, L. K. **Nonsymbiotic bacterial diazotrophs in cropfarming systems: can their potencial for plant growth promotion be better exploited?** Soil Biology & Biochemistry, Elmsford, v. 36, p. 1229-1244, 2006.

KHAMMAS, K. M., AGERON, E., GRIMONT, P. A. D.; KAISER, P. ***Azospirillum irakense sp. nov*., a nitrogen-fixing bacterium associated with rice roots and rhizosphere soil.** Res. Microbiol. v. 140, p. 679-693, 1989.,

KURTZ, S.; PHILLIPPY, A.; DELCHER, A. L.; SMOOT, M.; SHUMWAY, M.; ANTONESCU, C.; SALZBERG, S. L. **Versatile and open software for comparing large genomes**. Genome Biology. V. 5, R12, 2004.

LADHA, J. K.; TIROL-PADDRE, A.; DAROY, M. L. G.; PUNZALAN, G.; WATANABE, I. **The effect on N_2 fixation in vitro por bactérias diazotróficas endofíticas**. Pesquisa Agropecuária Brasileira. Brasília, v. 42, n. 10, p. 1459-1465, 2000.

LEMOS, M.; BASÍLIO, A.; CASANOVA, M. A. **Um estudo dos algoritmos de montagem de fragmentos de DNA**. PUC-RioInf.MCC, 2003.

LESK, A. **Introdução a Bioinformática**. 2º ed. Artmed, Porto Alegre, 2008.

LEWIS. S.; ASHBURNER, M.; REESE, M. G. **Annotating eukaryote genomes**. Current Opinion in Structural Biology, v. 10, p. 349-354, 2000.

MCKERNAN, K.; MARBLEHEAD, M. A. BLANCHARD, A.; MIDDLETON M. A.; KOTLER, L.; ALLSTON, M. A.; COSTA, G.; ESSEX, M. A. **Reagents, methods, and libraries for bead-based sequencing.** US patent application 20080003571, 2006.

MAGALHÃES, F. M., BALDANI, J. I., SOUTO, S. M., KUYKENDALL, J. R.; DOBEREINER, J. **A new acid-tolerant Azospirillum species**. Academia Brasileira Ciências. v. 55, p. 417-429, 1983.

MAXAM, A. M., GILBERT, W. **A new method for sequencing DNA**. Proc. Natl. Acad. Sci. v. 74, p 560-564, 1977.

MARTIN-DIDONET, C. C. G.; CHUBATSU, L. S.; SOUZA, E. M.; KLEINA, M.; REGO, F. G. M., RIGO, L. U.; YATES, M. G,; PEDROSA, F. O. **Genome structure of the genus Azospirillum**. Journal of Bacteriology, v. 182, n. 14, p. 4113-4116, 2000.

MATHWORKS Disponível em: http://www.mathworks.com/products/matlab/index.html Acesso em: 07/01/2012

MEHNAZ, S.; WESELOWSKI, B.; LAZAROVITS, G.; **Azospirillum zeae sp. Nov., a diazotrophic bacterium isolated from rhizosphere soil of Zea mays**, International Journal of Systematic and Evolutionary Microbiology, Great Britain, v. 57, p. 2805-2809, 2007.

MEHNAZ, S.; WESELOWSKI, B., LAZAROVITS, G. ***Azospirillum canadense* sp. nov., a nitrogen-fixing bacterium isolated from corn rhizosphere.** Int J Syst Evol Microbiol, v. 57, p. 620-624, 2007.

MEIDANIS, J.; SETÚBAL, J. C. **Uma introdução à biologia computacional**. IX Escola de Computação. Recife, 1994.

MOENS, S., MICHIELS, K., KEIJERS, V., VAN LEUVEN, F.; VANDERLEYDEN, J. **Cloning, sequencing, and phenotypic analysis of laf1, encoding the flagellin of the lateral flagella of *Azospirillum brasilense Sp7*.** Journal of. Bacteriology. v. 177, p. 5419-5426, 1995.

MIR, L. **Genômica**. Ed. Atheneu, São Paulo, 2004.

OKON, Y. e VANDERLEYDEN, J. Root-associated **Azospirillum species can stimulate plants**. ASM News, v. 63, n. 7, p. 366-370, 1997.

OVERBEEK, R.; BARTELS, D.; VONSTEIN, V.; MEYER, F. **Annotation of bacterial and archael genomes: improving accuracy and consistency**. Chem Rev. v. 107, n. 8, p. 3431-3447, 2007.

PASSALACQUA, K. D.; VARADARAJAN, A.; ONDOV, B. D.; OKOU, D. T.; ZWICK, M. E.; BERGMAN, N. H. **Structure and complexity of a bacterial transcriptome**. Journal of Bacteriology. v. 34, n. 3, p. e22, 2009.

PENG, G.; WANG, H.; ZHANG, G.; HOU, W., LIU, Y., WANG, E. T.; TAN, Z. **Azopirillum melinis sp. nov., a group of diazotrophs isolated from tropical molasses grass**. Int J Syst Evol Microbiol v. 56, p. 1263-1271, 2006.

PROSDOCIMI, F. **Introdução a bioinformática**. Biotecnologia Tecnologia Ciência & Desenvolvimento. Brasília, 2007 Disponível em: http://biotec.icb.ufmg.br/chicopros Acesso em: 10/01/12.

RAMOS, R. T. J. **Desenvolvimento de uma "*suite*" de aplicativos computacionais para suporte à montagem de genomas bacterianos a partir de leituras curtas**. Programa de Pós-graduação em Genética de Biologia Molecular / UFPA. 2011.

RAMOS, R.T.J.; CERNEIRO, A.R; BAUMBACH, J.; AZEVEDO, V.; SCHNEIDER, M.P.C. AND SILVA, A. **Analysis of quality raw data of second generation sequencers with quality assessment software**. BMC Research Notes, v. 4, p.130, 2011.

REINHOLD, B., HUREK, T., FENDRIK, I., POT, B., GILLIS, M., KERSTERS, K., THIELEMANS, S. DE LEY, J. ***Azospirillum halopraeferens* sp. nov., a nitrogen-fixing organism associated with roots of Kallar Grass (Leptochloa fusca (L.) Kunth).** Int. J. Syst. Bacteriol. v. 37, p. 43-51, 1987.

ROESCH, L. F.; CAMARGO, F. O.; SELBACH, P. A.; SÁ, E. S. **Reinoculação de bactérias diazotróficas aumentando o crescimento de plantas de trigo.** Ciência Rural, Santa Maria, v 35, nº 5, p 1201-1204, 2005.

ROUZÉ, P.; PAVY, N.; ROMBAUTS, S. **Genome annotation: which tools do we have for it?** Current Opinion in Structural Biology. v. 2, p. 90-95, 1999.

SANGER, F., AIR, G. M., BARRELL, B. G., BROWN, N. L., COULSON, A. R., FIDDES, J. C. **Nucleotide sequence of bacteriophage phiX174 DNA**. Nature*, v. 265 p.* 687-695, 1977.

SHENDURE, J.; JI, H. **Next-generation DNA sequencing**. Nature Biotechnology, v. 26, n. 10, p. 1135-1145, 2008.

SMITH, A. D.; XUAN, Z.; ZANG, M. Q. **Using quality score and longer reads improves accurancy of Solexa read mapping.** BMC bioinformatics. v. 9, p. 128, 2008.

SUMMER, M. E. Crop responses to **Azospirillum inoculation.** Advances in Soil Sciences, New York, v 12, p 54-123, 1990.

STEIN, L. **Genome annotation: from sequence to biology**. Nature reviews Genetics, v. 2, p. 493-503. 2001.

STERKY, F.; LUNDEBERG, J. **Sequence analysis of genes and genomes**. Journal of Biotechnology. v. 76, p. 1-31, 2000.

STEENHOUD, O.; VANDERLEYDEN, J. ***Azospirillum*, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic.** Biochemical Microbiology. Review. v. 24, p. 487-506, 2000.

SHUSTER, S. C. **Next-generation sequencing transform today's biology**. Nat. Method, v. 5, p. 16-18, 2008.

TARRAND, J.J.,KRIEG, N.R.; DOBEREINER, J. **A taxonomic study of the Spirillum lipoferum group, with descriptions of a new genus, Azospirillum gen. nov. and two species, *Azospirillum lipoferum* (Beijerinck) comb. nov. and *Azospirillum brasilense sp. nov***.. Can. J. Microbiol. v. 24, p. 967-980, 1978.

TURCATTI, G.; ROMIEU, A.; FEDURCO, M. TAIRI, A. P. **A new class of cleavable fluorescent nucleotide: synthesis and optimization as reversible terminators for DNA sequencing by synthesis.** Nucleic acids research, v. 36, n. 4, p. e25, 2008.

ZERBINO, D., & BIRNEY, E. Velvet: Algorithms **for de novo short read assembly using de Bruijn graphs**. Genome Research*, v. 18*, p. 821-829, 2008.

ZHULIN, I.B. and ARMITAGE, J.P. **Motility, chemokinesis, and methylation independent chemotaxis in Azospirillum brasilense**. J. Bacteriol. v. 175, p. 952-958, 1993.

XIE, C. H.; YOKOTA, A. ***Azospirillum oryzae* sp. nov., a nitrogen-fixing bacterium isolated from the roots of the rice plant Oryza sativa**. Int J Syst Evol Microbiol, v. 55, p, 1435-1438, 2005.

WISNIEWSKI-DYÉ, F.; BORZIAK, K.; KHALSA-MOYERS, G.; ALEXANDRE, G.; SUKHARNIKOV, L. O. WUICHET, K.; HURST, G. B.; MCDONALD, W. H.; ROBERTSON, J. S.; BARBE, V.; CALTEAU, A.; ROUY, Z.; MARGENOT, S.; PRIGENT-COMBARET, C.; NORMAND, M. B.; SIGUIER, P.; DESSAUX, Y.; ELMERICH, C.; CONDEMINE, G.; GANISAN, K.; KENNEDY, I.; PETERSON, A. H.; GONZÁLEZ, V.; MAVINGUI, P.; ZHULIN, I. B. **Azospirillum genomes reveal transition of bacteria from aquatic to terrestrial environments**. PLOS Genetics. V. 7, 2011.